THE *Journal* OF
Experimental Education

Volume 44, Number 4 Summer 1976

In this issue:

THE JOURNAL OF EXPERIMENTAL EDUCATION

THE *Journal* OF EXPERIMENTAL EDUCATION

Volume 44, Number 4 Summer 1976

CONTENTS

The Journal of Experimental Education is published four times a year by HELDREF publications, 4000 Albemarle St., N.W., Washington, D.C. 20016. Annual subscription rates are $12.50 for institutions and $10 for individuals, plus $3 postage for all subscriptions outside the United States and Canada. Single copies $3. Second class postage paid at Washington, D.C. All business correspondence should be sent to this address. Claims concerning missing issues made within 6 months will be serviced free of charge. Send all manuscripts to Prof. John Schmid, Department of Research and Statistical Methodology, University of Northern Colorado, Greeley, CO 80631.

Arvil S. Barr, Founder

EDITOR AND PUBLISHER © 1932-1962

(The Journal of Experimental Education is indexed/abstracted in Abstr. S. W., CSPA, Current Contents, Ed. Adm. Abst., Educ. Ind., Soc. of Ed. Abst., Current Index to Journals in Education, Language and Language Behavior Abst.)

DIMENSIONS OF TEACHING EFFECTIVENESS: A STUDENT PERSPECTIVE

BETTY J. HASLETT
University of Delaware

ABSTRACT

Forty-one semantic differential scales measuring the concept of a good teacher were factor analyzed to assess the general, underlying judgmental dimensions which students use in evaluating effectiveness in teaching. Both high school and college students judged teachers on the dimensions of student/teacher rapport, communicative style, instructional style, and stimulation. A personalization factor, measuring the teacher's ability to personalize and make relevant class materials, was also found among college students and conceptually distinguished them from the high school students. In addition to class level differences, sex differences across students were also investigated.

THE SEARCH FOR EFFECTIVENESS in teaching has been a focal point of educational research for decades. Recently, the increasing pressure for academic accountability has focused more attention on the evaluation of instruction. One technique for evaluating instructional effectiveness is the use of student ratings of instructors and courses. In their comprehensive review of the literature on evaluations, Costin, Greenough, and Menges (6) concluded that student ratings of instruction were valid and reliable in assessing various criteria of instructional effectiveness.

In studies involving student ratings of instructor and course, an increasing number have utilized factor analytic techniques to uncover the underlying judgmental dimensions involved in evaluating teaching behavior. These general factors identify basic components of the teaching process and are therefore of importance in training and evaluating teachers. Meredith (15) concluded that research on the dimensionality of student ratings, although not extensive, supported either a two-factor model focusing on instructor empathy and instructional competency or a six-factor model isolating the factors of general course attitude, instructional method, student interest and attention, course content, instructor characteristics, and specific procedures. The research of Finkbeiner, Lathrop, and Schuerger (9) supported a multifactor model of course attitudes: they found five factors interpreted as general course attitude; attitude toward examinations; attitude toward method; instructor/student rapport; and attitude toward workload. A review of seven-factor analytic studies among college students by Cashin (3) revealed several factors common to all the studies. These common factors were: course organization; workload or difficulty level of courses; student-teacher rapport and interaction; general teaching skills; instructor impact; stimulation and interest; and grading and evaluation methods.

However, these studies suffer from several limitations. First, factor structures generated from these studies rely on actual, observed conduct of the teacher in the classroom, thus sampling a particular student's experience in a particular class at a specific time. Such dimensions suggest factors operating in particular educational contexts, not general underlying components of teaching behavior. As Ryans (17) has observed, little reliable information is available regarding good teaching: one of the major reasons contributing to this situation, he concluded, was "the lack of any clear understanding of the various patterns of behavior that characterize teachers in general" (17:1).

A second limitation of factor analytic studies is that such studies have been done exclusively with college students. Several researchers (17, 19) found that class level significantly influenced students' judgments of teaching effectiveness. These findings suggest that studies assessing the dimensionality of student ratings of courses and instructors should be extended to include students at other academic levels. This researcher is aware of only one factor analytic study, conducted by Smalzreid and Remmers (18), which utilized high school students. Smalzreid and Remmers used the Purdue Rating Scale for Instructors and found two factors, an empathy trait (which appears to be similar to student/teacher interaction factors in college factor analytic studies) and a professional maturity trait (which appears to be similar to teaching method and instructor competency factors in college factor analytic studies). This study, however, used only a selected subsample of items from the Purdue Rating Scale for Instructors.

The general question to be answered, then, is whether certain types of students view good teachers in ways that differ from the perspective of other students. More specifically, the present research attempted to (*a*) characterize the general dimensions that underlie students' assessment of teaching behavior; and (*b*) assess what changes, if any, occur in students' general evaluations of teaching effectiveness as a function of the educational experience (either high school or collegiate experience) and sex of the student. It was hypothesized that the basic underlying dimensions of assessing effective teaching would be the same for both high school and college students. However, the relative contribution toward effective teaching that each dimension made would differ for the two groups.

Method

Subjects

Ss were 667 high school students and 219 college students. There were 282 males and 385 females in the high school subject pool, and 98 females and 121 males in the college subject pool.

Design

The general abstract concept of a "good teacher" was selected to be measured because it assesses one's attitude toward teaching in general and one's judgment of the various criteria used to evaluate teaching. A semantic differential was constructed to evaluate the concept of a good teacher. Fifteen high school students wrote essays describing the best teacher as well as the worst teacher they had had. From these essays, a set of adjectives was abstracted and a series of bipolar adjectival scales for the concept constructed. Sixty scales were used to evaluate the concept of a good teacher. A pilot study was conducted among 90 college students to evaluate each potential scale item.

In addition to evaluating their instructor using the test instrument, these students were asked to delete any scales they thought irrelevant or unimportant, and to add scales reflecting qualities they believed important for effective teaching if not already included in the test instrument. On the basis of the deletions and additions suggested by students in the pilot study and the experimenter's deletion of redundant or unreliable scale items, 41 scales were selected for inclusion in the final form of the test instrument.

Procedure

The instrument was administered to the high school students in their English classes. Students were told that this research was to aid in constructing instruments for the evaluation of instruction. Students were also told to rate good teaching in general and not the instructor of any class they were taking or had taken in the past. Four college classes from the state university were selected to participate in the study and represented a cross section of students from the humanities, social, and natural sciences.

Directions for using the semantic differential scales were printed on the top of the instrument scales and were read to the *Ss*. *Ss* were encouraged to answer every scale. The *E* or the research assistant worked through several examples to further illustrate use of the scales.

Data Analysis

The data analysis done in this study reported the results of factor analyses using a principal components solution with an orthogonal, varimax rotation done on the scales measuring the concept of a good teacher for both high school and college students. Unities were inserted for communalities, and the point at which substantive increases in the cumulative proportion of variance were no longer being made served as a cut-off point in the iteration of factors. A principal components solution was used to generate orthogonal, independent factor scores for each *S*. Multivariate analyses of variance were also done to test for significant differences in student scores across the scale items as a function of student sex and educational experience.

Results

High School Students' Judgmental Dimensions in Evaluating Good Teachers

After a number of varimax rotations were done, it was found that a four-factor solution accounting for 41% of the total variance was the most meaningful interpretation of the data. The first factor, *student/teacher rapport*, was characterized by the qualities of trustworthiness, fairness, cooperativeness, and openness, and accounted for 54% of the variance explained by the factors. Factor II, *communicative style*, accounted for 20% of the factor variance and was measured by qualities such as ease or difficulty in understanding the teacher's remarks, being comfortable in the classroom, being interesting and available for student consultation. *Instructional style*, the third general evaluative dimension, accounted for 14% of the factor variance and reflected teaching skills such as general organization, knowledge of the material, experience, and intelligence. The last factor, *stimulation*, reflected how challenging, strict, and difficult a teacher was, and accounted for 12% of the factor variance. (The factor structure of effective teaching ratings among high school students is presented in Table 1.)

College Students' Judgmental Dimensions in Evaluating Good Teachers

After a number of varimax rotations, a five-factor solution accounting for 43% of the variance was found to be the best interpretation of the data. The first factor, *student/teacher rapport*, accounted for 50% of the factor variance and was measured by scales such as responsiveness, fairness, trustworthiness, and concern for students. Factor II, *instructional style*, accounting for 15% of factor variance, was

Table 1.–High School Students' Factor Structure for the Concept of a Good Teacher*

I		II		III		IV	
Student/Teacher Rapport		Communicative Style		Instructional Style		Stimulation	
clear	.740	interesting	.626	experienced	.665	challenging	.738
trustworthy	.738	admits errors	.594	organized	.646	strictness	.671
fair	.680	open-minded	.541	intelligent	.598	demanding	.651
presents other views	.670	comfortable in class	.519	sticks to the point	.573		
concerned	.635	available	.518	knowledge of material	.515		
cooperative	.573	originality	.486	precise	.463		
humor	.569	easy to understand	.448				
responsive	.545						
easy to talk to	.492						
competent	.486						

*–Factor structure and variable loadings as obtained from SSPS factor analytic program using varimax rotation

Table 2.–College Students' Factor Structure for the Concept of a Good Teacher*

I		II		III		IV		V	
Student/Teacher Rapport		Instructional Style		Communicative Style		Stimulation		Personalization	
fair	.734	intelligent	.671	admits errors	.584	demanding	.695	uses classtime effectively	.731
clear	.680	sticks to the point	.643	humor	.557	critical	.687	personalizes material	.702
trustworthy	.672	knowledge of material	.557	informal	.527	challenging	.614	sensitive	.468
concerned	.662			originality	.466			open-minded	.442
responsive	.657	organized	.489	congenial	.456				
decisive	.524	experienced	.487	interesting	.444				
presents other views	.479	appears comfortable in class	.449						
competent	.478	interesting	.446						
flexible	.474	energetic	.426						

*–Factor structure and variable loadings as obtained from SSPS factor analytic program using varimax rotation

measured by scales such as organization, intelligence, experience, knowledge of the material, and interestingness. The third factor, *communicative style*, accounted for 13% of the factor variance and was measured by scales such as congeniality, sense of humor, willingness to admit mistakes, and informality. Factor IV, *stimulation*, accounted for 12% of the variance and measured a teacher's ability to be challenging and stimulating. The last factor, *personalization*, accounted for 10% of the variance and reflected the teacher's ability to add a personalized, human quality to his teaching. This factor was measured by scales such as ability to personalize and make class material relevant, and the ability to be sensitive and open-minded. (The factor structure for college students' evaluations of teaching effectiveness is set forth in Table 2.)

Sex Differences in Evaluation of Good Teaching

Since the sex of students participating in the experiment was known, the data were analyzed to see what differences, if any, existed between the evaluations of males and females with regard to teaching effectiveness. Factor scores for each S were generated and a multivariate analysis of variance done to assess differences in judgments between males and females. Among the college students, only Factor I, *student/teacher rapport*, showed any significant differences in judgments between males and females: females found good teachers to be significantly more responsive, trustworthy, concerned, and so forth than did males, $F(1, 213) = 8.42, p < .01$. Among high school students, only Factor III, *instructional style*, showed

significant differences in evaluations of effective teaching between males and females, $F(1, 665) = 4.51, p < .05$. Females judged good teachers to be significantly more organized, experienced, intelligent, and knowledgeable than did males.

Student Sex and Academic Environment: Their Effects across Individual Semantic Differential Scales

In order to analyze the effect of student sex and academic environment more closely, a multivariate analysis of variance was done on the semantic differential scales to explore their effects upon students' evaluations of the characteristics of effective teachers. Significant differences in students' evaluations were found from the main effects of student sex and academic environment, $F(41, 844) = 2.53$ and 5.63, respectively; $p < .001$. Fifteen of 41 scales, or 36% of the scales, showed significant differences between male and female in their assessment of various criteria measuring teaching effectiveness. Overall, females rated good teachers more highly than males on 90% of the scales.

With regard to the influence of academic environment, college and high school students differed significantly from one another in their judgments on 21 of 41 scales, or 51% of the scales. College students judged effective teachers more positively than did high school students on 36 of 41 scales, or 87% of the scales. Table 3 presents the overall means for all students categorized by sex and academic environment. There was s significant interaction between sex X academic environment on only one scale, "agressiveness," with college males judging teachers as being most agressive, followed by high school males, college females, and high school females, $F = 6.07, p < .01$.

Discussion

Both groups of students, high school as well as college, appeared to have very similar judgmental dimensions with regard to evaluating good teaching. Both factor analyses revealed a fundamental duality in judgment: teachers were judged as professionals (e. g., in their role as teachers) and as individuals (e. g., as unique personalities). Many researchers (7, 17) have observed that educators need to understand what the basic behavioral components of teaching are before one can begin to evaluate and measure teaching effectiveness, as well as train effective teachers.

Both high school and college students judged teaching on the dimensions of student/teacher rapport, communicative style, instructional style, and difficulty or stimulation level. Although the variance accounted for by these evaluative dimensions varied for each group, and there were some differences in the variables loading on these factors for each group, the similarity of the four factors for both high school and college students suggests that these four evaluative dimensions are basic, fundamental skills needed for good teaching, regardless of subject matter taught, level of students taught, teacher personality, or a number of other variables.

Previous factor analytic studies of student ratings (20) have found evaluative dimensions similar to the judgmental dimensions found in the present investigation. Various dimensions of effective teaching may interact with one another to produce particular teaching styles that are most effective for certain educational contexts, students, or subject areas: not all dimensions will be equally stressed for differing contexts, students, or subject areas. Yet the basic similarity of the judgments found in this study and in previous investigations suggests that these four facets of teaching behavior would undoubtedly be reflected in any particular teaching strategy.

Past research on the qualities of good or ideal college teachers (1, 4, 10, 11) found that qualities such as thorough knowledge of the subject matter, clarity, open-mindedness, and the ability to be interesting were characteristics associated with effective teaching. The present study found these qualities to be valued by both high school and college students: other studies have shown somewhat similar rankings and agreement among students, faculty, and administrators (2, 8, 16). Table 4 gives a comparison of the rank order of the ten most valued characteristics for effective teachers by high school and college students in the present study (selected by rank-ordering the scales from highest to lowest mean—the higher the mean, the more valued the scale), as well as the ten most valued qualities found in the Bousfield, Clinton, and Perry studies (1, 4, 16).

The personalization factor, not found in previous factor analytic studies among college students, may reflect a change in American society and higher education over the last decade: a sense of growing alienation and depersonalization as documented in Toffler's *Future Shock* and Reich's *The Greening of America.* The educational experience in high school is more personal and individualized since classes are usually small and students and teachers typically know each other to some extent, whereas the college experience is more impersonal and isolated due to large classes, the greater number of students enrolled, and the impersonal contact between teacher and student, and often times, between student and student.

Given the substantial changes in ideas and attitudes students undergo during their college years (20), students undoubtedly feel the need for personal attention, relevancy, and meaning in their interaction with others in the academic community during this period of personal growth and change. Since for many students teachers provide the only source of extended personal contact within the academic structure (excluding, of course, personal friendships that develop among students), it is not surprising that college students believe that teachers in particular ought to personalize their teaching and make their subject areas meaningful and relevant for students. Such an attitude may in part be documented by the growth of programs

Table 3.—Effect of Academic Environment and Student Sex on Student Attitudes toward Good Teachers*

Scales	High School		College		p value Sex effect	p value Environmental effect
	Males	Females	Males	Females		
Agressive/Unagressive	4.97	4.76	4.93	5.19	NS	.010
Easy to talk to/ Not easy to talk to	6.29	6.28	6.31	6.57	NS	NS
Flexible/Inflexible	5.39	5.30	5.68	5.86	NS	.001
Personalizes material/ Material presented abstractly	4.52	4.67	5.10	5.24	NS	.001
Uses classtime effectively/ Does not use classtime effectively	5.49	5.73	6.14	6.28	.010	.001
Open-minded/Narrow-minded	6.26	6.36	6.32	6.61	NS	NS
Sensitive/Insensitive	4.95	4.92	5.51	5.85	NS	.001
Informal/Formal	4.95	5.18	5.28	5.29	NS	NS
Sense of Humor/Humorless	6.08	5.96	5.96	6.13	NS	NS
Shows originality/Does not show originality	5.93	5.97	6.04	6.21	NS	NS
Challenging/Easy	4.85	5.00	5.20	5.52	NS	.001
Strict/Lenient	3.45	3.60	3.72	3.82	NS	.010
Unprejudiced/Prejudiced	6.09	6.06	6.23	6.31	NS	.030
Intelligent/Unintelligent	4.43	4.33	4.85	5.01	NS	.001
Precise/Imprecise	6.25	6.47	6.40	6.52	.010	NS
Sticks to the point/ Digresses	4.93	5.21	5.39	5.40	.020	.010
Critical/Uncritical	6.10	6.22	6.15	6.47	NS	NS
Fair/Unfair	5.77	6.14	5.73	6.13	.001	NS
Energetic/Unenergetic	6.51	6.62	6.52	6.67	.050	NS
Trustworthy/Untrustworthy	5.90	6.07	6.33	6.52	.030	.001
Experienced/Inexperienced	5.22	5.13	5.89	5.84	NS	.001
Congenial/Uncongenial	6.09	6.27	6.24	6.47	.010	.040
Knows Material/Does not know material	5.99	6.19	6.26	6.52	.010	.001
Concerned/Indifferent	6.42	6.40	6.47	6.71	NS	.020
Competent/Incompetent	6.08	6.20	6.00	6.27	NS	NS
Easy to understand/Difficult to understand	5.62	6.01	5.97	6.13	.001	NS
Cooperative/Uncooperative	6.07	6.02	6.23	6.35	NS	.004
Demanding/Undemanding	5.41	5.27	5.81	5.92	NS	.001
Appears comfortable in class/ Does not appear comfortable in class	4.58	4.70	4.58	4.57	NS	NS
Extroverted/Introverted	4.37	4.19	4.95	4.63	.010	.001
Considers student opinions/ Does not consider student opinions	6.34	6.55	6.40	6.66	.001	NS
Does not show favoritism/ Shows favoritism	5.74	5.89	6.05	6.28	.040	.001
Interesting/Uninteresting	6.36	6.48	6.36	6.57	NS	NS
Decisive/Indecisive	5.58	5.82	5.97	6.16	.007	.002
Responsive/Unresponsive	5.45	5.56	5.76	5.56	NS	NS
Presents different viewpoints/Presents only one viewpoint	6.28	6.37	6.64	6.71	NS	.001
Available/Unavailable	5.96	6.18	6.24	6.47	.003	.001
Clear/Unclear	5.85	5.93	6.41	6.58	NS	.001
Organized/Unorganized	6.27	6.38	6.21	6.33	NS	NS
Admits mistakes/Does not admit mistakes	6.33	6.45	6.35	6.52	NS	NS
Important to me/Unimportant to me	5.47	5.85	5.41	5.81	.001	NS

*p values determined by conducting F tests with 1 degree of freedom in the numerator and 842 degrees in the denominator. Scales are consistently arranged with the left-hand member of the bipolar adjectives representing the more positive judgment of that value (and having high numbers represent that end of the continuum) and the right-hand member representing the more negative evaluation of a particular characteristic (represented by lower numbers). For some scales, such as "Formal/Informal," where a judgment of which attribute is preferred is not clear, this scale arrangement has been decided arbitrarily by the experimenter.

Table 4.–Characteristics of Good Teachers*

Clinton (1930)	Bousfield (1940)	Perry (1971)	High School *Ss*	College *Ss*
Knowledge of subject	Fairness	Well prepared for class	Clarity	Knowledge of subject
Pleasing personality	Mastery of subject	Sincere interest in subject	Trustworthiness	Interesting
Neatness in work and appearance	Interesting style of presentation	Knowledge of subject	Challenging	Clarity of presentation
Fairness	Well organized	Effective teaching methods	Fairness	Fairness
Kind, sympathetic	Clarity of presentation	Tests for understanding	Strictness	Competency
Sense of humor	Interest in students	Fairness	Presents others' views	Trustworthiness
Interest in profession	Helpfulness	Effective communication	Experienced	Open-mindedness
Interesting style of presentation	Ability to direct discussion	Encourages independent thought	Organized	Admits mistakes
Alertness and broad-mindedness	Sincerity	Logical organization of course	Concern for students	Responsiveness
Knowledge of methods	Keen intellect	Motivates students	Interesting	Available to students

*–In rank order of their importance in each study

like Women's Studies, Black Studies, and Criminal Justice, which reflect student and societal demands for relevancy and social impact.

Since the current investigation involved a relatively small college population and sampled a limited geographic locale, these results can only be suggestive of a new dimension of teaching effectiveness. Additional research will be necessary to explore further the personalization factor and examine its significance for theories of instruction, learning, and personal growth.

The sex differences found in the factor structures for high school and college students centered around two factors: female college students rated an effective teacher more highly on the rapport dimension than did males, and female high school students rated effective teachers' instructional style more highly than did male high school students. The difference among the high school students was interpreted as a reflection of the more negative, critical contacts that male high school students have with their teachers (12): males, as a result of this negative interaction, will have a less positive opinion of teachers in general.

The female college students' higher rating of an effective teacher's rapport was interpreted as a result of their greater response to warmth and openness on the part of teachers (14). When examining individual semantic differential scales for sex differences across scales, it was found that females rated good teachers significantly higher on scales measuring the teacher's warmth and openness. This was thought to reflect not only the female college student's greater responsiveness to a teacher's warmth but also the more positive contacts that high school females have with their teachers (12). Generally, females rated good teachers more highly than did males.

College students rated effective teachers more positively than did high school students. Examining these differences across individual semantic differential scales, it was found that college students rated effective teachers significantly higher on three general parameters: a teacher's competency (measured by scales such as knowledge of material, intelligence, and experience); a teacher's level of difficulty (reflected by scales such as being challenging and demanding); and a teacher's responsiveness (measured by scales such as extroversion, concern, availability, and sensitivity). These differences were interpreted as reflecting the greater degree of educational difficulty in college–students, as well as teachers, are expected to be more knowledgeable and competent in general, as well as responsive to more vigorous intellectual demands. The personalization factor found among collegians was also thought to influence their ratings of a teacher's responsiveness. In general, college students rated good teachers more positively than did high school students; this may generally reflect the higher value college students place upon education since such students have elected to further their education.

As Trent and Cohen (20) concluded in their review of research on teaching in higher education, there is a need for research on societal expectations that define a teacher's role. Such knowledge is necessary for effective teaching, adequate teacher training, and accurate teaching evaluation. The Levinthal, Lansky, and Andrews study (13) found that there is an interaction between a student's concept of an ideal teacher and that student's ratings of actual instructors.

Before one can properly evaluate a student's ratings of instruction it is necessary to understand the educational values and role expectations of teachers which guide the student's ratings (17). The current research was a step in the direction of assessing some of these values and role expectations among students of differing sex and educational experiences.

REFERENCES

1. Bousfield, W.A., "Students' Ratings of Qualities Considered Desirable in College Professors," *School and Society*, 51: 253-256, 1940.
2. Brewer, R.; and Brewer, M.B., "Relative Importance of Ten Qualities for College Teaching Determined by Pair Comparisons," *Journal of Educational Research*, 63:243-246, 1970.
3. Cashin, W.E., *Student Ratings of Teaching*, Research Memorandum 74-1, Academic Planning and Evaluation, University of Delaware, 1973.
4. Clinton, R.J., "Qualities College Students Desire in College Instructors," *School and Society*, 32:702, 1930.
5. Coffman, W.E., "Determining Students' Concepts of Effective Teaching from Their Ratings of Instructors," *Journal of Educational Psychology*, 45:277-286, 1954.
6. Costin, F.; Greenough, W.; and Menges, R., "Student Ratings of College Teaching: Reliability, Validity and Usefulness," *Review of Educational Research*, 41:511-535, 1971.
7. Dick, W., "Course Attitude Questionnaire: Its Development, Uses and Research Results," as revised by D. Stickell, Office of Examination Services, Pennsylvania State University, September 1967.
8. Drucker, A.; and Remmers, H.H., "Do Alumni and Students Differ in their Attitudes toward Instructors?," *Purdue University Studies in Higher Education*, 70:62-64, 1950.
9. Finkbeiner, C.; Lathrop, J.; and Schueger, J., "Course and Instructor Evaluations: Some Dimensions of a Questionnaire," *Journal of Experimental Psychology*, 64:159-163, 1973.
10. French, G.M., "College Students' Concept of Effective Teaching Determined by an Analysis of Teacher Ratings," *Dissertation Abstracts*, 17:1380-1381, 1957.
11. Gadzella, B.M., "College Student Views and Ratings of an Ideal Professor," *College and University*, 44:89-96, 1968.
12. Good, T.L.; Sikes, J.N.; and Brophy, J.E., "Effects of Teacher Sex and Student Sex on Classroom Interaction," *Journal of Educational Psychology*, 65:74-87, 1973.
13. Levinthal, C.F.; Lansky, L.M.; and Andrews, O.E., "Student Evaluations of Teacher Behaviors as Estimations of Real-Ideal Discrepancies–A Critique of Teacher Rating Methods," *Journal of Educational Psychology*, 62:104-109, 1971.
14. McKeachie, W.J.; Lin, Y.; and Mann, W., "Student Ratings of Teacher Effectiveness: Validity Studies," *American Educational Research Journal*, 8: 435-445, 1971.
15. Meredith, G.M., "Dimensions of Faculty-Course Evaluations," *Journal of Psychology*, 73:27-32, 1969.
16. Perry, R.R., "Evaluation of Teaching Behavior Seeks to Measure Effectiveness," *College and University Business*, 68:18-22, 1971.
17. Ryans, D., *Characteristics of Teachers*, American Council on Education, Washington, D.C., 1960.
18. Smalzreid, N.T.; and Remmers, H.H., "A Factor Analysis of the Purdue Rating Scale for Instructors," *Journal of Educational Psychology*, 34:363-367, 1943.
19. Tolor, A., "Evalution of Perceived Teacher Effectiveness," *Journal of Educational Psychology*, 64:98-104, 1973.
20. Trent, J.W.; and Cohen, A.M., "Research on Teaching in Higher Education," in R.M.W. Travers (ed.), *The Second Handbook of Research on Teaching*, Rand McNally, Chicago, 1973.

AN EMPIRICAL ANALYSIS OF THE INSTRUCTIONAL EFFECTIVENESS IN VISUALIZED INSTRUCTION

THOMAS C. ARNOLD
FRANCIS M. DWYER
The Pennsylvania State University

ABSTRACT

The purpose of this study was to investigate the relative effectiveness of specific media attributes on student performance on criterion tests measuring different levels of understanding. Specifically, it attempted to identify which of two levels of stimulus explicitness in visuals was most effective in facilitating student achievement on criterion tests of knowledge, comprehension, and total understanding for students possessing two different levels of entering behavior. One hundred seventy-one subjects participated in the study. The two-way ANOVA procedure was utilized to investigate the existence of interaction between entering behavior and level of stimulus explicitness. Results indicated that a significant relationship existed between entering behavior and performance on post-criterion tests; no relationship existed between stimulus explicitness and achievement on the criterion tests; and insignificant interactions were found to exist between entering behavior and instructional treatment.

A NUMBER OF EDUCATIONAL RESEARCHERS (2, 4, 6, 7) commenting on teaching effectiveness have indicated that current media research has not incorporated instructional techniques based on sound instructional and/or psychological research. A great many of the studies seem to be primarily concerned with conducting evaluative comparisons to support the use of one form of media in preference to another, while providing little insight concerning the effectiveness of attributes inherent to a particular medium. Recently educators have encouraged the development of research designs which would not only evaluate the relative effectiveness of different media but would also identify instructional strategies by which given types of learners would achieve optimally. One of the theoretical orientations which has emerged as a result of this trend was designed by Salomon (6) and is used in this study. His theory of stimulus explicitness attempts to present an understanding of how the use of media influences learning. Since this theory was influential on the design of this study, a brief synopsis is warranted.

Theoretical Orientation

It is Salomon's belief that for stimuli to be effective in learning they must affect mental processes in the learner relevant to the task being learned. The stimulus explicitness theory assumes that one of the most fundamental functions of visual stimuli is to inform, that is, to reduce uncertainty and thus increase the learner's probability of achieving a correct response relevant to the learning task. He further suggests that the instructional effectiveness of a given type of visual stimulus in reducing uncertainty is contingent upon the prior existence of aroused uncertainty in the individual.

Different types of visual materials contain varying amounts of realistic detail which, in turn, can be considered to represent varying degrees of stimulus explicitness. For example, if in a learning situation the individual does not experience any uncertainty, his behavior might depict what is normally called the boredom syndrome—daydreaming, doodling, etc. However, if the stimulus materials used in the instructional situation generate some uncertainty, the learner may be motivated to search for additional information in order to reduce this uncertainty. If too much uncertainty is introduced into the learning situation, it may cause the learner to react negatively towards the simulus materials and reject the purposes for which they were originally designed.

This assumption finds support in the theory and from information theorists (3, 5, 6) who report that as the amount of information in the stimulus increases, the uncertainty generated by the stimulus decreases. Figure 1 illustrates this relationship between uncertainty and stimulus explicitness.

In this continuum the amount of uncertainty in a stimulus is a function of the amount of information conveyed by the stimulus. As the amount of information in a visual increases, the uncertainty generated by the visual decreases. In terms of information theory this could be interpreted to mean that visuals possessing higher degrees of explicitness should have a greater potential for reducing entering uncertainty, thus increasing the probability that a greater amount of learning will occur.

Figure 2 illustrates the relationship between Salomon's theory and information theory. This diagram graphically

depicts the relationship between learning probability and the amount of uncertainty generated by visuals containing different amounts of stimulus explicitness.

This figure illustrates that the probability of learning increases as the uncertainty in the stimulus decreases due to an increase in explicitness of the stimulus. However, since Salomon predicts that this curve will vary depending on the learner's prior cognitive experience, he suggests that the direction of the learning curve will be dependent on the entering uncertainty of the individual. This projected relationship between entering behavior, learning probability, stimulus explicitness, and stimulus uncertainty is shown in Figure 3.

The purpose of this study was to evaluate the predictability of Salomon's stimulus explicitness theory by investigating the instructional effectiveness of two types of visual stimuli each possessing different degrees of stimulus explicitness. Specifically, the purposes of this study were to: (*a*) explore the research potential of the stimulus explicitness theory as a model for guiding research on visualized instruction; and (*b*) determine the instructional effectiveness of visual materials possessing different degrees of stimulus explicitness and also their effect on students possessing different entering behaviors.

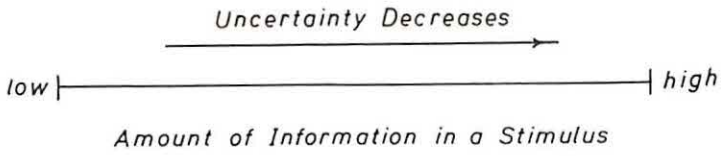


Figure 1.–An uncertainty continuum

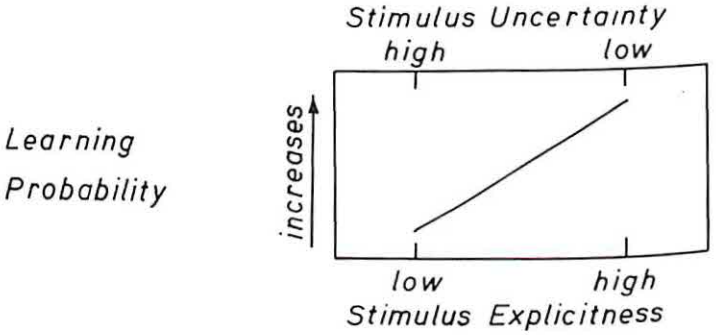


Figure 2.–Relationship between learning probability and the explicitness or uncertainty in a stimulus

Method

Materials

The materials for this investigation consisted of two sets of instructional programs designed in textbook format. The printed subject matter transmitted via these instructional packages was held constant, with each package consisting of 37 pages. Each page contained a 2½ inch by 3½ inch illustration of the human heart that was designed to complement the printed content material on that page.

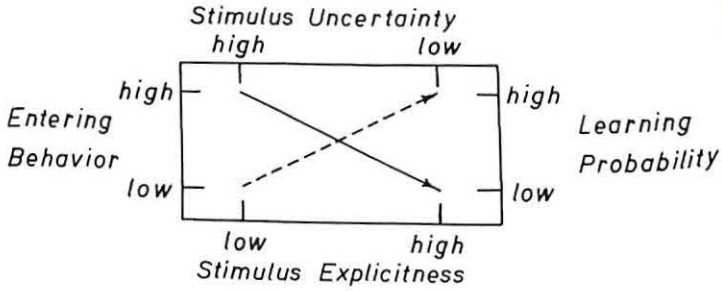


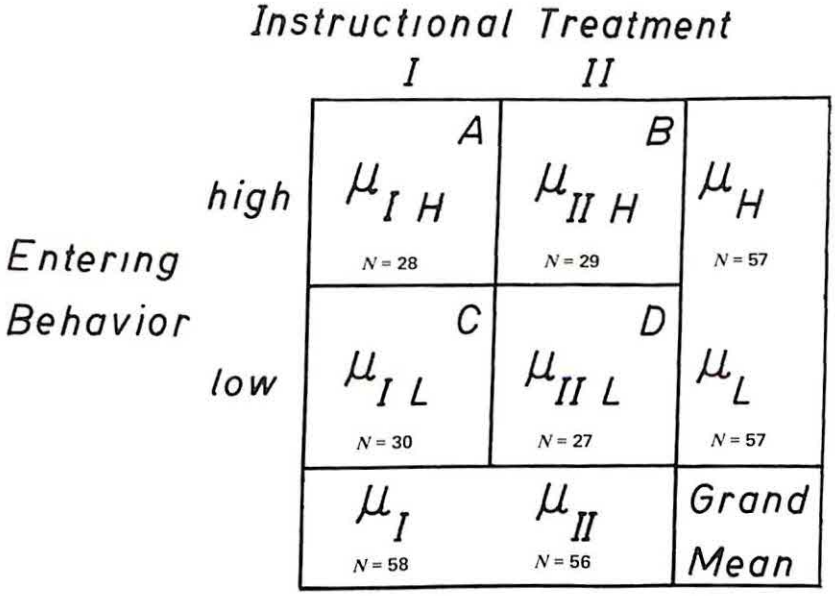
Figure 3.–Relationship between entering behavior, learning probability, and the explicitness or uncertainty in a stimulus

Treatment Groups

One hundred seventy-one college students enrolled in the Instructional Media 411 course at the Pennsylvania State University participated in this study. This course provides the orientation and competencies recommended by the State of Pennsylvania as necessary requirements for teaching certification. The findings of this study may well be generalized to students majoring in education.

Students were assigned to one of three entering behavior groups as a result of their performance on a pre-test employed for this purpose. Members of each entering behaviro group were then randomly assigned to one of two treatment groups. These treatment groups received identical written presentations; however, each of the two groups received their own respective type of visual illustrations containing one of two degrees of stimulus explicitness (uncertainty). Detailed, shaded drawings (high stimulus explicitiness/low uncertainty) of the human heart were employed in the instructional packages to complement the printed instruction.

Figure 4 shows the experimental design for this study. In order to insure significant differences between entering behavior groups and to provide reliability in assignment to the different groups, only those students identified as achieving high and low entering behavior on the content pre-test were used in this study. The statistical procedure employed for this purpose involved the establishment of confidence limits about a student's obtained score, and resulted in the probability of .95 that students used in this study were, in fact, correctly assigned to the proper treatment groups.



A: high entering behavior—simple pictures
B: high entering behavior—detailed pictures
C: low entering behavior—simple pictures
D: low entering behavior—detailed pictures

Figure 4.—The experimental design

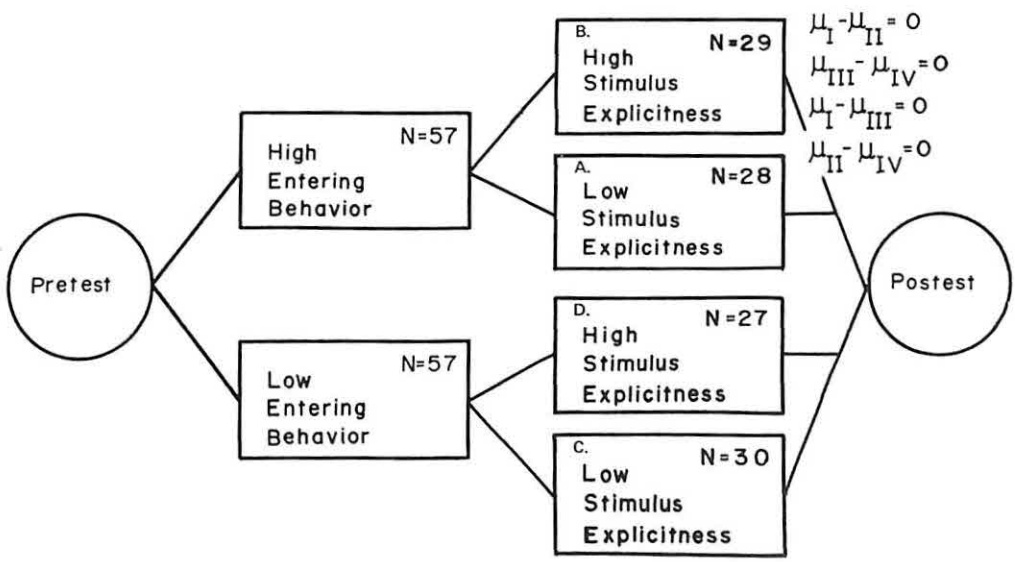


Figure 5.—The experimental design with derivative null hypothesis

Criterion Measures

Each student received a pre-test before receiving his respective treatment. In addition, each student also received a 44-item critical test which consisted of two subtests, each designed to measure a specific level of cognitive ability. Achievement scores on the two criterial measures were the dependent variables, and the degree of stimulus explicitness in the visualized treatments was the independent variable.

On the pre-test (K-R 20, r = .64) students were given a diagram of the heart and a series of questions asking them to identify specific parts of the heart. After completion of the instructional booklet, Ss received the 44-item post criterial test which consisted of the subtest of knowledge (K-R 20, r = .80) and the subtest of comprehension (K-R 20, r = .65). Scores from the two subtests were combined to provide a measure of total understanding (K-R 20, r = .85).

Data Analysis

A two-factor analysis of variance was the statistical model used to analyze the data obtained. Difference were considered significant at the .05 level. The two-way analysis of variance model was also used to investigate the interaction effect between entering behavior and instructional treatment. The data obtained in this study are listed as collected for specific sub-problems. Because the use of the 2 X 2 factorial design often provides for the testing of numerous statistical hypotheses, Figure 5 has been provided to aid the reader in visualizing the derivation of the null hypotheses. This figure represents the experimental design that was followed for each of the three sub-problems. Table 1 illustrates the means and standard deviations for students in each treatment group on each criterion measure.

Unweighted means were used in the ANOVA procedures since the study was exploratory in nature and there was no reason to expect that one treatment would be more effective than the other. The ANOVA used was adapted to handle unequal *N*'s; furthermore, the number of students in each cell was not greatly disproportionate so that the variance would be seriously affected. Bartlett's test for homogeneity was on the pre-test scores, and in no case did the observed values reach the critical value for a .05 level test. Thus, it appeared that the students receiving the different treatment could, in fact, be considered to have been drawn randomly from populations with common variance.

Sub-Problem # 1

Which of two levels of stimulus explicitness is most effective for learners identified as having either high or low entering behavior as measured by a test of knowledge of terminology on achievement in learning as measured by the total test of understanding? (See Table 2.)

Table 2.–Results of the Two-Factor ANOVES of the Data from Entering Behavior and Treatment as Measured by the Total Test of Understanding

Source of variation	*df*	*ms*	*F*	*p*
Entering behavior	1	763.87	16.40*	0.00
Treatment	1	31.25	.67	0.42
Entering behavior X treatment	1	1.09	.02	0.88
Within groups	110	46.59		

*Significant at 0.05

Sub-Problem # 2

Which of two levels of stimulus explicitness is most effective for learners identified as having either high or low entering behavior on achievement in learning as measured by the specific criterion test–the test of knowledge? (See Table 3.)

Table 3.–Results of the Two-Factor ANOVES of the Data from Entering Behavior and Treatment as Measured by the Test of Knowledge

Source of variation	*df*	*ms*	*F*	*p*
Entering behavior	1	296.14	17.22*	0.00
Treatment	1	21.53	1.25	0.27
Entering behavior X treatment	1	0.003	0.00	0.99
Within groups	110	17.20		

*Significant at 0.05

Sub-Problem # 3

Which of two levels of stimulus explicitness is most effective for learners identified as having either high or low entering behavior on achievement in learning as measured by the specific criterion test–the test of comprehension? (See Table 4.)

Table 1.–Means and Standard Deviations for Students in Each Treatment Group on Each Criterion Measure.

Criterion measures	High stimulus explicitness (*N* = 29)		Low stimulus explicitness (*N* = 28)		High stimulus explicitness (*N* = 27)		Low stimulus explicitness (*N* = 30)	
	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*
Total test of understanding	28.79	6.88	30.04	6.14	23.81	7.21	24.67	7.01
Test of knowledge	15.62	4.42	16.50	3.61	12.41	4.21	13.27	4.29
Test of comprehension	13.17	3.43	13.54	2.90	11.41	3.49	11.40	3.36

Table 4.—Results of the Two-Factor ANOVES of the Data from Entering Behavior and Treatment as Measured by the Test of Comprehension

Source of variation	*df*	*ms*	*F*	*p*
Entering behavior	1	108.77	9.96*	0.00
Treatment	1	0.90	0.08	0.77
Entering behavior X treatment	1	0.98	0.09	0.77
Within groups	110	10.92		

*Significant at 0.05

Results

Three conclusions were derived from the data obtained in this study:

1. There was a significant relationship between entering behavior and performance on criterion tests. Those students whose prior experience and knowledge of the content material were high performed more effectively than those with low entering behavior regardless of the type of visual illustration they received.
2. No significant relationship was found to exist between the level of stimulus explicitness and achievement on the criterion tests. This was interpreted as meaning that visuals possessing either of the stimulus explicitness levels were equally effective in improving achievement of identical objectives for each of the entering behavior groups.
3. No significant interactions were found to exist between entering behavior and type of visualization.

Summary and Discussion

The purpose of this study was to investigate the relative effectiveness of specific media attributes on student performance on criterial tests measuring different levels of understanding. More specifically, this study was designed to gather data to ascertain which of two levels of stimulus explicitness in a series of visuals provided for the most effective instruction as measured by achievement on criterial tests of knowledge, comprehension, and total understanding for students possessing two different levels of entering behavior.

The theoretical rationale for this study was Salomon's theory of stimulation. Thus, while gathering data for purposes of investigating the above problem, it was also the objective of this study to investigate that portion of Salomon's theory that pertained to the role of visuals in the instructional process. The assumption of the theory's having application to this study implied that research in involving the stimulus-explicitness dimension of visuals should provide educators with the means to determine how effective a specific visual is. That is, research associated with this dimension should provide data to determine how much cognitive activity occurs as a result of exposure to a specific kind of stimulus presentation by a particular learner.

These two objectives are related in the following manner. The stimulus-explicitness dimension in a visual is a function of the amount of information conveyed by the visual and the ability of the information to reduce the learner's uncertainty, thereby increasing his probability of learning whatever message that visual was designed to convey. The theory postulates that data gathered from research associated with this dimension should indicate the degree that cognitive processes are activated after exposure to this stimulus attribute. Stated in reference to this study, if the stimulus-explicitness property of a visual affects the probability of reducing aroused uncertainty, then different levels of stimulus explicitness should activate different cognitive processes. If this were the case, then at Bloom's cognitive levels of terminology, and comprehension, one would expect that different visuals would differ in their effectiveness to improve learning measured at these cognitive levels. In other words, according to the assumptions of the theory one would expect high entering behavior students receiving less stimulus explicitness in visuals to perform as effectively at the same cognitive levels as low entering behavior students who received visuals possessing a higher level of stimulus explicitness. This assumption is predicated on the concept that high entering behavior students initially experience lower levels of uncertainty and consequently require less explicitness to achieve an equal probability for successful performance on achievement tests measuring learning at different levels of understanding. Conversely, if given the higher levels of explicitness in visuals, high entering behavior students should experience a reduction in their probability to attain high achievement because they are not receiving the optimum form of instruction.

Applying these expectations to low entering behavior students, one could expect those receiving the optimum form of instruction to perform better on the achievement tests for each cognitive level than those receiving the less optimal instruction. That is, low entering behavior students receiving visuals possessing low levels of stimulus explicitness would experience a reduction in their probability to attain as high an achievement score than if they had received the more optimal instruction possessing visuals with the higher level of stimulus explicitness.

The data collected for this study did not support these assumptions in Salomon's theory of stimulation. More specifically, an analysis of the findings failed to produce any significant interactions between the stimulus-explicitness level in visuals and the entering behavior of students on the criterion tasks.

Conclusions

Three major conclusions can be made with some degree of confidence concerning the experimental problem.

1. For students identified as having either high or low entering behavior there were significant differences between their mean scores on each of the post-criterion tests. Though one could argue that these differences are valid because the groups possessed significant differences in entering behavior relevant to the instructional material, other interpretations seem to be warranted. Analyzing these significant differences in terms of Salomon's theory of stimulation, it would appear that the two treatments were not effective in increasing a student's probability for learning in the direction predicted by the theory. That is to say, students receiving the different instructional treatments did not demonstrate different performance levels. These data seem to contest Salomon's theory of stimulation and simultaneously support Dwyer's research (1) which contends that reality can be edited for instructional purposes.

2. In regard to the two treatment groups, there were no significant differences between the mean scores of the first treatment group (those receiving visuals having low stimulus explicitness) as measured from each of the post-criterion tests. This could be interpreted to mean that visuals possessing either of the two levels of stimulus explicitness were equally effective in enhancing achievement on identical objectives for each of the entering behavior groups

3. In regard to the presence of interaction effects between entering behavior and instructional treatment, the data showed that there were no systematic effects on performance on the criterion tests due to a combination of a particular entering behavior with a particular method of instruction. This conclusion also contests that segment of Salomon's theory postulating that this form of interaction should occur.

REFERENCES

1. Dwyer, F.M., *A Guide for Improving Visualized Instruction*, Pennsylvania State University Learning Services, State College, Pa., 1972.
2. Funk, C.E., "Instructional Efficiency with Biological Objects in a Task Requiring Dichotomous Identification Keying Techniques," unpublished doctoral dissertation, Pennsylvania State Universtiy, 1971.
3. Garner, W.R., *Uncertainty and Structure as Psychological Concepts*, Wiley, New York, 1962.
4. Lutz, J.E., "Instructional Efficiency with Three-dimensional Objects in Tasks Requiring Identification Keying Techniques," unpublished doctoral thesis, Pennsylvania State University, 1970.
5. Moser, G.W., *The Use of Information Theory to Study Human Learning*, symposium paper presented at the 1973 Annual Meeting of the National Association for Research in Science Teaching, March 29, 1973.
6. Salomon, G., "What Does It Do to Johnny? A Cognitive-Functionalistic View of Research on Media," in G. Salomon and R.E. Snow (eds.), *ViewPoints*, Bulletin of the School of Education, Indiana University, Spring 1970.
7. Travers, R.M.W., ed., *Second Handbook of Research on Teaching*, Rand McNally, Chicago, 1973.

COVARIANCE AND DISCRIMINANT ANALYSIS

CARL J. HUBERTY
University of Georgia

ABSTRACT

The formal equivalence between tests of the effect of deleting variables in a multiple response setting based on distances and based on multivariate analysis of covariance (MANCOVA) is shown for the two-group case. Implications of this result for the interpretation of results of multiple-group analyses are discussed, as well as how MANCOVA may be useful for ordering and selecting variables that contribute to group discriminations.

THE BASIS FOR MUCH STUDY in the domain of multivariate statistical analysis has been Fisher's (3) introduction of a classification technique which dealt with the problem of assigning objects, on which there are p correlated measures available, to one of two well-defined populations. The formal equivalence of this technique to that of multiple regression analysis with a dichotomous criterion variable is well known. This equivalence has been

shown to extend to the related problems of variable deletion and interpretation of discriminant functions. Collier (1) showed that tests used in deleting variables in regression analysis and in classification are equivalent, while Huberty (7) showed that predictor variables may be equivalently ordered (with respect to contribution to discrimination) by univariate F-ratios and by estimates of predictor versus discriminant function correlations. The related problems of deleting variables and interpreting discriminant functions may be considered a part of the important problem of selecting the best subset of predictors for optimal classification. These partial solutions do provide a basis for variable selection procedures; attempts have been made to generalize the predictor versus discriminant function correlation idea to the k-group case.

The primary purpose of the present note is to show that a test of the equality of the distance between two populations when based on p predictors and the distance when based on q predictors ($q < p$) is equivalent to a test of the equality of the two population mean vectors (or centroids) using multivariate analysis of covariance (MANCOVA) with $p - q$ variates and q covariates. The first-mentioned test is that of Rao (10: 482), and may be expressed as a test of

$$H_o : \Delta_p^2 = \Delta_q^2$$

where Δ_p^2 is the generalized distance function between (the centroids of) two populations based on p variables. Mahalanobis's (8) distance (squared) between the two populations as estimated from the sample on the basis of the p variables is

$$D_p^2 = \mathbf{d}' \left(\frac{W}{N_1 + N_2 - 2} \right)^{-1} \mathbf{d} \qquad [1]$$

where N_j = size of sample (or group) j,

$\mathbf{d}$ = $(p \times 1)$ vector of differences between means on the variables for the two samples, and

W = $(p \times p)$ within-groups sums of squares and cross products (SSCP) matrix.

The test statistic used is:

$$\frac{N_1 N_2 (D_p^2 - D_q^2)}{(N_1 + N_2)(N_1 + N_2 - 2) + N_1 N_2 D_q^2} \cdot \frac{N_1 + N_2 - p - 1}{p - q} \qquad [2]$$

which may be referred to the F distribution with $p - q$ and $N_1 + N_2 - p - 1$ degrees of freedom under H_o with the usual assumptions.

The statistic used in testing

$$H_o : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$$

using MANCOVA, as given by Rulon and Brooks (11:94), is:

$$\frac{N_1 N_2 (N_1 + N_2 - 2) D_{p|q}^2 \; / \; N_1 + N_2 - q - 2)}{[(N_1 + N_2)(N_1 + N_2 - 2) + N_1 N_2 D_p^2]} \cdot \frac{N_1 + N_2 - p - 1}{p - q} \qquad [3]$$

where $D_{p|q}^2$ is the adjusted distance (as will subsequently be explicitly defined). The referrent distribution for this statistic is the same as that for [2]. (It is noted here that [3] is a transformation of Hoteling's T^2 statistic with covariance adjustment.) The purpose of this note, then is to show that [2] and [3] are equal.

To begin with some notation, one may consider the $(p \times p)$ supermatrix of SSCP:

$$W = \left[\begin{array}{c|c} W_{11} & W_{12} \\ \hline W_{21} & W_{22} \end{array} \right]$$

where

W_{11} = $([p - q] \times [p - q])$ within-groups SSCP matrix of the $p - q$ variates,

W_{22} = $(q \times q)$ within-groups SSCP matrix of the q covariates,

W_{12} = $([p - q] \times q)$ within-groups cross products of the $p - q$ variates and q covariates, and

$W_{21} = W_{12}'$

Furthermore, let

$$M = W_{11} - W_{12} W_{22}^{-1} W_{21} \qquad [4]$$

which is a $(p - q) \times (p - q)$ matrix.

The distance (squared) $D_{p|q}^2$ in [3] as defined by Rulon and Brooks (11:94), using the above notation, is

$${}_a\mathbf{d}' \left(\frac{M}{N_1 + N_2 - q - 2} \right)^{-1} {}_a\mathbf{d}$$

where ${}_a\mathbf{d} = (p - q \times 1)$ vector of differences between adjusted variable means. Now, for the $p - q$ variates and q covariates, the $(p - q \times q)$ matrix of regression weights is

$$B = W_{12} W_{22}^{-1}$$

Thus, ${}_a\mathbf{d} = \mathbf{d}_1 - W_{12} W_{22}^{-1} \mathbf{d}_2$

where

$\mathbf{d}_1 = (p - q \times 1)$ vector of mean differences between the two groups on the $p - q$ variates, and

$\mathbf{d}_2 = (q \times 1)$ vector of mean differences between the two groups on the q covariates.

Since in general $(A/k)^{-1} = kA^{-1}$ where A is a nonsingular matrix and k is a scalar, the numerator of the fraction on the left of [3] may be expressed as

$$N_1 N_2 (N_1 + N_2 - 2)(\mathbf{d}_1 - W_{12} W_{22}^{-1} \mathbf{d}_2)' M^{-1} (\mathbf{d}_1 - W_{12} W_{22}^{-1} \mathbf{d}_2)$$

or

$$N_1 N_2 (N_1 + N_2 - 2)(\mathbf{d}_1' - \mathbf{d}_2' W_{22}^{-1} W_{21}) M^{-1} (\mathbf{d}_1 - W_{12} W_{22}^{-1} \mathbf{d}_2) \qquad [5]$$

To prove what we set out to prove, it is sufficient to show that the numerator of the fraction on the left of [2] is equal to [5].

Now, it may be noted that this numerator in [2] can be written as

$$N_1 N_2 (N_1 + N_2 - 2)[\mathbf{d}' W^{-1} \mathbf{d} - \mathbf{d}_2' W_{22}^{-1} \mathbf{d}_2]$$

By partitioning $\mathbf{d}'$ as $[\mathbf{d}_1' \vdots \mathbf{d}_2']$ and considering the inverse of a supermatrix (5:469), $\mathbf{d}' W^{-1} \mathbf{d}$ may be expressed as

$$[\mathbf{d}_1' \vdots \mathbf{d}_2'] \left[\begin{array}{c|c} M^{-1} & -M^{-1} W_{12} W_{22}^{-1} \\ \hline -W_{22}^{-1} W_{21} M^{-1} & W_{22}^{-1} + W_{22}^{-1} W_{21} M^{-1} W_{12} W_{22}^{-1} \end{array}\right] \left[\begin{array}{c} \mathbf{d}_1 \\ \hline \mathbf{d}_2 \end{array}\right]$$

where M is defined by [4]. Thus, we have that

$$\mathbf{d}' W^{-1} \mathbf{d} = [\mathbf{d}_1' M^{-1} - \mathbf{d}_2' W_{22}^{-1} W_{21} M^{-1} \vdots -\mathbf{d}_1' M^{-1} W_{12} W_{22}^{-1} + \mathbf{d}_2' W_{22}^{-1} W_{21} M^{-1} W_{12} W_{22}^{-1}] \left[\begin{array}{c} \mathbf{d}_1 \\ \hline \mathbf{d}_2 \end{array}\right]$$

$$= [\mathbf{d}_1 M^{-1} \mathbf{d}_1' - \mathbf{d}_2 W_{22} W_{21}^{-1} M^{-1} \mathbf{d}_1 - \mathbf{d}_1' M^{-1} W_{12} W_{22}^{-1} \mathbf{d}_2 + \mathbf{d}_2' W_{22}^{-1} W_{21} M^{-1} W_{12} W_{22}^{-1} \mathbf{d}_2]$$

And, therefore,

$$\mathbf{d} W^{-1} \mathbf{d} - \mathbf{d}_2' W_{22}^{-1} \mathbf{d}_2 = \mathbf{d}_1' M^{-1} \mathbf{d}_1 - \mathbf{d}_2' W_{22}^{-1} W_{21} M^{-1} \mathbf{d}_1 - \mathbf{d}_1' M^{-1} W_{12} W_{22}^{-1} \mathbf{d}_2 + \mathbf{d}_2' W_{22}^{-1} W_{21} M^{-1} W_{12} W_{22}^{-1} \mathbf{d}_2$$

which simplifies to the triple matrix product in [5]. In essence, then, what has been shown is that the total distance function (D_p^2) can be written as the sum of the distance function for the covariates (D_p^2) and the distance function for the main variates after adjustment for the covariates ($D_{p|q}^2$); i. e.,

$$D_p^2 = D_q^2 + D_{p|q}^2$$

In terms of multivariate tests of hypotheses, what has been shown is the formal equivalence of (*a*) a test of the equality of distances between (the centroids of) two populations before and after some of the original variables have been deleted, and (*b*) a test of the equality of two population centroids using a covariance analysis of the questionable variables with the significant variables as covariates. The common statistic yields an indicator of whether or not there remains any appreciable useful information in the data after the significant variables have been removed.

There is a limited practical implication as well as a theoretical implication of the proven result. It is not necessary that a researcher interested in deleting variables execute two runs of a discriminant analysis program to obtain D^2 values, and then calculate the statistic yielded by [2] to pass judgment on a subset of doubtful variables. What can be done is to employ a MANCOVA program, which should be available at most computer centers. (Repeated analyses may be carried out using different subsets.)

More important, however, is a potentially valuable theoretical implication of the above result which may be generalizable to situations where $k > 2$. This implication pertains to the problem of variable selection and interpretation in multivariate analysis. The following procedure is a suggestion made by Hall (4: 8-9):

1. Select the variable with the highest univariate F-ratio and rerun the analysis using this variable as a covariate.

2. From the results of the reanalysis (a MANCOVA), select from the remaining variables that variable which yields the highest univariate F-ratio. Assign it as an additional covariate, and reanalyze.

3. Continue steps 1 and 2 until the multivariate F-ratio shows no significant variation (say, $p < .10$) among the means of the remaining variables.

This procedure, which probably tends to overestimate the number of significant variables, could be modified by a stepwise procedure (4:9). Cramer and Bock (2: 607) indicate that such a use of MANCOVA to delete variables is implicit in Rao's work (10) as a generalization of the two-group case.

MANCOVA was used by Horton, Russell, and Moore (6) for selecting the most effective discriminators; the method is similar to that of Hall except that it involves more analyses to determine the significant variables. The procedure used begins with the smallest subset of variables (i. e., one) and then, after testing all possible combinations with the complementary subset, selects that combination which left the smallest residual, as indicated by the smallest value of a test criterion (a likelihood ratio statistic). If the value of the criterion of the selected subset is significant, an additional variable is included in the subset of variables to be retained using the same procedure as before. This cycle of operations is terminated when no significant residual between-group variance remains. In the 12-group situation of this study, five out of nine original variables were retained.

In addition to selecting a good subset of discriminators, the use of MANCOVA also provides the researcher with an ordering of the variables with respect to contribution to discrimination among the criterion groups. This may be helpful for interpretive reasons, especially in comparing results across studies involving similar variables. The same may be said when discriminant functions are obtained and interpreted following the rejection of an hypothesis—for main or interaction effects—in a factorial multivariate analysis of variance. In such a case, follow-up univariate analyses may be considered—with possible adjustments in nominal α-levels—and, assuming more than one discriminator is "significant," it may be well to carry out a MANCOVA in interpreting the additional contribution of succeeding single discriminators.

Some comments about the use of MANCOVA in variable selection may be made. First of all, no claim can be made that such a selection procedure would yield the best subset of the resulting size. What would be necessary, of course, is to examine all possible subsets of a given size. Secondly, a "forward deletion" or step-up procedure as described is open to criticism, in the sense that after a deletion the analysis is carried out on the "bad" variables, and it can be said that in so doing much of what is being analyzed may be "noise." Proponents of such a view would prefer a "backward deletion" or stepdown procedure in which the bad variables are discarded from the analysis at each step. This issue also arises with variable selection in multiple regression analysis; Mantel (9) points out other advantages of a variable selection scheme in which the variables are successively discarded one at a time from the original full set.

REFERENCES

1. Collier, R. O., "A Note on the Multiple Regression Technique for Deleting Variables in the Discriminant Function," *Journal of Experimental Education,* 31: 351-353, 1963.
2. Cramer, E. M.; and Bock, R. D., "Multivariate Analysis," *Review of Educational Research,* 36: 604-617, 1966.
3. Fisher, R. A., "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics,* 7: 179-188, 1936.
4. Hall, C. E., *Three Papers in Multivariate Analysis,* American Institutes for Research, Palo Alto, Calif., 1967.
5. Horst, P., *Matrix Algebra for Social Scientists,* Holt, Rinehart and Winston, New York, 1963.
6. Horton, I. F.; Russell, J. S.; and Moore, A. W., "Multivariate-Covariance and Canonical Analysis: A Method for Selecting the Most Effective Discriminators in a Multivariate Situation," *Biometrics,* 24: 845-858, 1968.
7. Huberty, C. J., "Regression Analysis and Two-Group Discriminant Analysis," *Journal of Experimental Education,* 41: 39-41, 1972.
8. Mahalanobis, P. C., "On the Generalized Distance in Statistics," *Proceedings of the National Institute of Science, India,* 12: 49-55, 1936.
9. Mantel, N., "Why Stepdown Procedures in Variable Selection?," *Technometrics,* 12: 621-625, 1970.
10. Rao, C. R., *Linear Statistical Inference and Its Applications,* Wiley, New York, 1965.
11. Rulon, P. J.; and Brooks, W. D., "On Statistical Tests of Group Differences," in D. K. Whitla (ed.), *Handbook of Measurement and Assessment in Behavioral Sciences,* Addison-Wesley, Reading, Mass., 1968.

THE STABILITY OF TEACHER RATINGS ON THE DEVEREUX ELEMENTARY SCHOOL BEHAVIOR RATING SCALE[1,2]

JANE D. WALLBROWN
Worthington, Ohio Public Schools

FRED H. WALLBROWN
Columbus, Ohio Public Schools

JOHN BLAHA
University of Detroit

ABSTRACT

The stability of Devereux ratings (eleven factors, three items) was investigated for a sample comprised of 67 subjects from the primary unit of a suburban school organized with open classrooms and vertical grouping. The time interval between ratings was approximately one year. The reliabilities differed substantially for the fourteen scores. For the eleven factors, the median *r* was .73, but individual *r*'s ranged from .82 for Comprehension through .49 for Achievement Anxiety. The reliabilities for the three single items compared favorably with those for the eleven factors even though the latter were obtained by summing the scores for several items.

AN INCREASING BODY of research data, for example (4-7), suggests that the Devereux Elementary School Behavior Rating Scale (3) constitutes a promising technique for understanding the learning and behavior patterns of children referred for psychological assessment. The Devereux is comprised of 47 behavioral items which are relevant to classroom achievement and/or adjustment. Three items—Unable to Change, Quits, and Slow Work—are scored singularly, but the remaining items are combined to obtain scores for the following behavioral factors: Classroom Disturbance; Impatience; Disrespect-Defiance; External Blame; Achievement Anxiety; External Reliance; Comprehension; Inattentive-Withdrawn; Irrelevant Responsiveness; Creative Initiative; and Need for Closeness to Teacher.

As noted by Littell (2:137), the items comprising the Devereux were selected and grouped with great care, but only limited reliability data were included. Specifically, Spivack and Swift (4:30-33) provided test-retest reliability estimates for a subsample of 128 children who were rated again one week after their initial rating. The reliability estimates obtained for this group were satisfactory, as indicated by a median reliability coefficient of .87 for all eleven factors and coefficients for individual factor scores which ranged from .85 through .91. As might be expected, the reliability coefficients of .72, .80, and .71 for the three individual items were somewhat lower than those for the eleven factors.

The usefulness of the Devereux as a technique for the diagnosis and remediation of learning and behavior disorders is contingent upon the stability of scores across relatively long periods of time. Consequently, the present study was designed to investigate the long-term reliability of Devereux scores for primary grade children in an open classroom setting.

Method

Subjects

The final sample was comprised of 67 children (35 boys, 32 girls) who were enrolled in the primary unit of a suburban elementary school from May 7, 1973, through May 17, 1974. The total first grade enrollment was 75 at the time of the initial ratings; however, the sample was reduced to 67 by the time the final ratings were obtained near the end of second grade.

The majority of the Ss were above-average in intelligence and were from upper-middle-class families as indicated by the father's occupational status and educational level. That is, the median educational level for fathers was college graduation, and most of them were employed in professional or managerial positions. The mean IQ for the final sample on the Cognitive Abilities Test—Primary I, Form 2 (8) was 114.7, with the *SD* of 13.8.

Data Collection

The school from which the sample was obtained is organized in accordance with the open classroom concept and also provides for vertical grouping at the primary level. With this arrangement, each of the eight teachers in the primary unit had a class comprised of first-, second-, and third-grade children. Consequently, eight teachers were involved in completing both the initial and final ratings for the sample.

The initial ratings were completed during a five-day period in the spring of 1973, and the second set of ratings was completed approximately one year later during a similar time period. The eight teachers who completed the initial ratings were provided with a one-hour training session one week before the rating period. A similar training period was held before the final rating period to familiarize the two new teachers with the rating procedure and to review the procedure for the six teachers involved in the initial ratings.

Data Analysis

Test-retest reliability estimates were obtained by computing product-moment correlations between the initial and final ratings for each of the eleven factors and three individual items. Standard errors of measurement were computed using the formula provided by Horst (1:294).

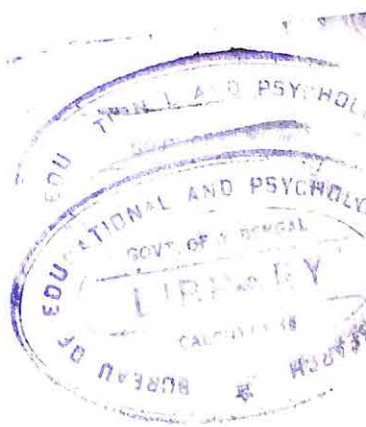
Results

The reliability coefficients for the fourteen Devereux scores (eleven factors and three items) are presented in Table 1, along with other relevant statistics. The standard error of measurement (*SEM*) for each score is included since this information is most important for the interpretation of Devereux ratings for individual children. The means and standard deviations for both the initial and final ratings are also included in Table 1. These data are important in that they establish the characteristics of the present sample and facilitate comparisons with other samples.

Examination of the *r*'s reported in Table 1 suggests that, for children in an open classroom setting, some of the Devereux scores are much more reliable than others. The median *r* for the eleven Devereux factors was .73, while the *r*'s for individual factor scores ranged from .49 through .86. The reliability estimate was highest for Comprehension with an *r* of .86; Disrespect-Defiance ranked second with an *r* of .82. The reliability estimate for Achievement Anxiety ($r = .49$) was substantially below that obtained for any other factor, but Impatience ($r = .62$) and Inattentive-Withdrawn ($r = .67$) were also relatively low. The median reliability estimate was the one obtained for External Reliance ($r = .73$). The reliability estimates for the other factors were closer to the median, with Irrelevant

Table 1.—Reliability Estimates and Other Statistics for Devereux Scores

Devereux Score	Statistic					
	Initial		Final			
	$\bar{X}$	*SD*	$\bar{X}$	*SD*	*r*	*SEM*
FACTOR:						
Classroom Disturbance	12.0	4.1	12.4	4.4	.75	3.0
Impatience	9.9	4.1	9.8	4.4	.62	3.7
Disrespect-Defiance	6.2	2.8	6.7	2.9	.82	1.7
External Blame	7.3	3.8	8.2	4.3	.71	3.1
Achievement Anxiety	9.7	4.1	9.7	3.9	.49	4.0
External Reliance	14.9	4.9	14.4	4.7	.73	3.5
Comprehension	13.9	3.4	14.8	3.4	.86	1.8
Inattentive-Withdrawn	8.2	4.4	7.8	3.6	.67	3.3
Irrelevant Responsiveness	6.7	2.7	7.0	2.7	.79	1.7
Creative Initiative	11.4	3.6	12.0	3.6	.78	2.4
Need for Closeness to Teacher	14.9	4.5	15.6	3.4	.70	3.1
EXTRA ITEMS:						
Unable to Change	2.3	1.5	2.4	1.5	.82	0.9
Quits	2.7	1.5	3.0	1.6	.74	1.1
Slow Work	2.7	2.0	2.8	1.8	.72	1.4



Responsiveness (r = .79); Creative Initiative (r = .78); and Classroom Disturbance (r = .75) somewhat above, and External Blame (r = .71) and Need for Closeness to Teacher (r = .70) slightly below.

The reliability estimates for the three individual items compare favorably with those for the eleven factors even though the factor scores are obtained by summing the ratings for several items. In fact, the reliabilities for Unable to Change (r = .82) and Quits (r = .74) were both above the median r for the eleven factors, and the reliability for Slow Work (r = .72) was only slightly below.

The *SEMs* reported in Table 1 provide concrete estimates of the extent to which raw scores on the Devereux tend to differ across a one-year time interval. One cannot determine from the *SEMs* to what extent score variability is the result of measurement errors and how much is attributable to systematic behavioral changes. However, the *SEMs* probably provide relatively accurate estimates of overall stability of Devereux scores across this time period. For example, the *SEM* for Classroom Disturbance is 3.0, which indicates that the chances are about 1 in 3 that one would obtain a random score change (increase or decrease) of 3 points or more. Similarly, the chances are about 1 in 20 of obtaining a score change of 6 points (2 *SEMs*) on the basis of random error. The *SEMs* for the remaining scores can be interpreted in a similar manner.

Discussion

Generally speaking, the reliability estimates discussed above are substantially lower than those reported by Spivack and Swift (4) for a one-week time interval. This overall pattern of differences is understandable since the test-retest interval for the present study was much longer, i.e., one year. Yet, determining the adequacy of long-term reliability estimates poses a real problem since, as noted by Horst (1:289), there is very little known about the relationship between reliability and the time interval between retests. Under these circumstances, one can only offer a tentative discussion of the reliability data and suggest that persons using the Devereux with individual children should consider scores individually in terms of their *SEMs*. Such an approach is necessary because the reliability estimates range from Comprehension with very good reliability, through Achievement Anxiety with very poor reliability.

In the final analysis, those individuals considering the Devereux for either group or individual use should evaluate its reliability in terms of their own unique situation and needs. Effective evaluation necessarily involves comparison of Devereux reliabilities with those for other rating scales, as well as consideration of the overall state of measurement in the area.

Those individuals interested in the reliability of the Devereux should also bear in mind the nature of the present sample and the specific type of curricular arrangement existing in the school where the ratings were obtained. Strictly speaking, the extent to which these reliability estimates can be generalized beyond the present sample is an empirical matter to be resolved through future research. However, it seems reasonable to surmise that these estimates would be most applicable to bright Ss from suburban schools with an open classroom arrangement. In contrast, one would not necessarily expect these estimates to describe the reliability of the Devereux for Ss in a traditional classroom setting.

NOTES

1. Appreciation is due Mr. Ronald Hopper, principal of Worthington Hills School, and the teachers in the primary unit for their full cooperation during the conduct of the study.

2. Reprint requests should be sent to Dr. Jane D. Wallbrown, Worthington City Schools, 55 East Stafford Avenue, Worthington, Ohio, 43085.

REFERENCES

1. Horst, P., *Psychological Measurement and Prediction*, Brooks-Cole, Belmont, Calif., 1968.
2. Littell, W., "Review of the Devereux Elementary School Behavior Rating Scale by G. Spivack & M. Swifts," in O.K. Buros (ed.), *The Seventh Mental Measurement Yearbook* (Vol. I), Gryphon Press, Highland Park, N.J., 1972.
3. Spivack, G.; and Swift, M., "The Devereux Elementary School Behavior Rating Scale: A Study of the Nature and Organization of Achievement-related Disturbed Classroom Behavior," *Journal of Special Education*, 1:71-90, 1966.
4. Spivack, G.; and Swift, M., *The Devereux Elementary School Behavior Rating Scale Manual*, Devereux Foundation, Devon, Pa., 1967.
5. Spivack, G.; Swift, M.; and Prewitt, J., "Syndromes of Disturbed Classroom Behavior: A Behavioral Diagnostic System for Elementary Schools," *Journal of Special Education*, 5:269-292, 1971.
6. Swift, M.; and Spivack, G., "The Assessment of Achievement-related Classroom Behavior," *Journal of Special Education*, 2:137-153, 1968.
7. Swift, M.; and Spivack, G., "Clarifying the Relationship between Academic Success and Overt Classroom Behavior," *Exceptional Children*, 36:99-104, 1969.
8. Thorndike, R.; Hagen, E.; and Lorge, I., *Cognitive Abilities Test*, Houghton Mifflin, Boston, 1968.

CREATIVITY TRAINING IN ELEMENTARY SCHOOLS IN BRAZIL

EUNICE ALENCAR
University of Brasilia

JOHN F. FELDHUSEN
FRED W. WIDLAK
Purdue University

ABSTRACT

The effects of the Purdue Creative Thinking Program (PCTP) on the creative abilities of elementary school children in an underdeveloped country were evaluated in an experiment with 578 Brazilian fourth- and fifth-graders. At each grade level, twelve classes were assigned to each of two treatment conditions (PCTP with reinforcement and PCTP without reinforcement) and a control group which had no exposure to PCTP. Pre- and post-testing with the Torrance Tests of Creative Thinking (TTCT) yielded twelve creativity measures. Using a 3 × 2 × 2 (treatment by sex by grade level) analysis of covariance, the creativity training was found to be effective, but reinforcement of pupil performance appeared to have a decremental effect.

WHILE THERE HAS BEEN a great deal of research on creativity in the United States (16) and a number of other countries (17), few studies have been conducted on the creative abilities of children from underdeveloped countries. Evidence from a number of sources (4, 10, 18) indicates that all children possess some creative potential. However, it is essential that the home and school provide conditions and instruction to help children realize their full potential. Schools in the United States devote much attention to creativity. Creativity should also be stressed in the schools of less well-developed nations. Through the development of creative thinking abilities in their children, these nations can make more rapid progress in the next generation.

In order to facilitate the development of creativity, a number of methods and programs have been designed. Two of the most widely researched and evaluated programs for elementary school children are the Productive Thinking Program (PTP), developed by Covington, Crutchfield, and Davies (1), and the Purdue Creative Thinking Program (PCTP), developed by Feldhusen et al. (6). Both programs are designed to strengthen cognitive skills which are central to the creative process and to provide experiences in creative thinking. Davis (3) reviewed methods and programs for teaching creative thinking and concluded that there are a number of successful methods and programs for teaching creative thinking.

The main purpose of this study was to determine whether the creative abilities of Brazilian children of elementary school age could be increased through the use of the PCTP (6).

PCTP involves 28 audiotapes and a set of three or four exercises for each tape. The taped program consists of a 3- to 4-minute presentation designed to teach a principle or idea for improving creative thinking, and an 8- to 10-minute story about a famous American pioneer, statesman, inventor, or researcher. The exercises contain printed directions and problems, or questions, which are designed to provide verbal and nonverbal practice in originality, flexibility, fluency, and elaboration in thinking.

Previous research with the program (5, 6, 8, 15) suggested that it is effective in improving creative thinking skills as measured by the Torrance Tests of Creative Thinking. Feldhusen et al. (5) studied the effects of the PCTP on the creative thinking abilities of children from third, fourth, and fifth grade, and found substantial gains in creative thinking abilities after 28 weeks of training. In a second study, Feldhusen and his co-workers (6) evaluated the effectiveness of the different parts of the PCTP in a sample of fourth-, fifth-, and sixth-grade children. Although the program or its components were not effective at all grade levels or for all criterion variables, there was considerable evidence for the overall effectiveness of the PCTP. Speedie, Treffinger, and Feldhusen (15) studied the long-range effects of the PCTP with a sample of upper elementary children and found results quite similar to those of Feldhusen et al. (6). Similar results were obtained by Shively et al. (14) in a study comparing the effectiveness of the Productive Thinking Program (1) and the PCTP in a sample of fifth-grade children. Feldhusen, Speedie, and Treffinger (8) summarized research involving the PCTP

Table 1.–Significant F Ratios for the Analyses of Covariance of Creativity Test Scores *

Sampling Unit	Source of Variation	*df*	Figural (L = Lines, PC = Picture Completion) Fluency L	Fluency PC	Flexibility L	Flexibility PC	Originality L	Originality PC
Individual students	Treatment	2	13.49 (411.56)	(0.43)	10.22 (124.18)	(0.62)	19.22 (1261.45)	25.56
	Sex	1	8.51 (259.58)	(1.60)	(14.35)	(2.62)	(13.01)	(0.03)
	Grade	1	(17.23)	17.72 (22.13)	(2.25)	(7.50)	(8.64)	(47.81)
	Residual	531	(30.49)	(1.24)	(12.14)	(1.94)	(65.62)	(9.87)
Classes	Treatment	2	6.38 (511.29)	(0.25)	6.00 (143.56)	(0.17)	11.67 (1369.14)	6.15 (26.01)
	Classes with -in treatments	20	2.76 (80.07)	3.22 (3.89)	2.04 (23.92)	(2.27)	(117.35)	(4.23)
	Subjects within classes	520	(29.02)	(1.21)	(11.70)	(2.01)	(63.90)	(10.30)

* $p < .01$
Mean squares in parentheses

Table 1.–Significant F Ratios for the Analyses of Covariance of Creativity Test Scores (cont.) *

Sampling Unit	Source of Variation	*df*	Verbal (U = Unusual Uses, PI = Product Improvement) Fluency U	Fluency PI	Flexibility U	Flexibility PI	Originality U	Originality PI
Individual students	Treatment	2	8.88 (572.54)	(72.01)	21.49 (176.79)	(0.25)	19.36 (413.76)	(62.72)
	Sex	1	9.30 (600.05)	(7.54)	(3.30)	(13.72)	(5.21)	(52.87)
	Grade	1	(6.16)	(141.97)	7.25 (59.63)	6.93 (34.70)	(0.90)	(4.93)
	Residual	531	(64.46)	(30.71)	(8.22)	(5.00)	(21.36)	(21.85)
Classes	Treatment	2	8.97 (821.01)	(43.36)	12.66 (182.95)	(0.20)	10.71 (473.96)	5.47 (72.68)
	Classes with -in treatments	20	(91.57)	(15.44)	(14.45)	(2.45)	2.14 (44.14)	(13.28)
	Subjects within classes	520	(64.95)	(31.57)	(8.22)	(5.19)	(20.58)	(22.38)

* $p < .01$
Mean squares in parentheses

and concluded that it is an effective program for grades three to six.

Previous research (7, 11) using reinforcement to encourage thinking indicated that written verbal comments on children's creative productions would be motivating to the children and would increase their fluency and originality. However, the stress placed on the avoidance of evaluation in creative thinking by Osborn (12) raised some doubts about the possible effects of this variable on creative thinking.

In the present study, fourteen of the twenty-eight stories of the PCTP and the corresponding exercises were used with a sample of children in Brazil. The choice of the fourteen dramatized stories was based on their relationship to the program of history and social studies in Brazilian schools. The programs were translated into Portugese by the first author.

Method

Sample

A total of 578 fourth- and fifth-grade children from 24 classes in both private and public elementary schools in Brasilia, Brazil, participated in the study. There were twelve fourth-grade and twelve fifth-grade classes, with eight classes assigned to each of two treatment conditions (program with reinforcement of the pupils' performance on the creativity exercises and program with no reinforcement of the pupils' performance on the creative exercises) and eight classes assigned to the control group condition.

Procedure

Before instruction began, two verbal sub-tests (Unusual Uses and Product Improvement) and two figural sub-tests (Circles and Picture Completion) of the Torrance Tests of Creative Thinking (TTCT), Form B (16), were administered as pre-tests to all pupils in both the experimental and control groups. The tests were translated into Portugese by the first author. The instructional material was then administered to the experimental groups by the teacher once a week for fourteen consecutive weeks. The teachers were taught how to use the material by the first author. In administering the program, the teacher read the introduction and the story to the children since tape players were not available. The pupils then worked on the printed exercises. In one experimental condition (program with reinforcement), the children's completed exercises were evaluated by the experimenter. She wrote encouraging comments on their papers intended to reinforce fluency and elaboration (e.g., *very good; good; good, but try harder; try harder*), and then gave them back to the children. Pupils in the other experimental condition (program with no reinforcement) received no reinforcement. Pupils in the control group received no creativity training.

At the conclusion of the series of fourteen programs, two verbal (Unusual Uses and Product Improvement) sub-tests and two figural (Lines and Picture Completion) sub-tests of the TTCT, Form A, were administered as post-tests to all pupils in the experimental and control groups.

A 3 × 2 × 2 (treatment by sex by grade level) analysis of covariance was used to analyze pupil performance on each of the twelve creativity measures: figural fluency, flexibility, and originality for the sub-tests of Lines and Picture Completion; and verbal fluency, flexibility, and originality for the sub-tests of Unusual Uses and Product Improvement. Previous research by the authors indicated that the creativity sub-tests were task-specific and should be analyzed separately. The covariates for the divergent thinking measures were the respective TTCT pre-test measures. Post hoc individual comparisons between adjusted means were made for significant effects using the Newman-Keuls procedure. Further analyses of covariance were carried out to analyze the effect of treatment using the class as the sampling unit. Alpha was set at .01 for all tests of significance.

Results

The results of the analyses of covariance are summarized in Table 1. Using the individual subject as the sampling unit, a consistent finding across all dependent variables was that no interaction effect reached statistical significance. The main effect of treatment was significant for all three creativity dimensions of fluency, flexibility, and originality for the Lines and Unusual Uses sub-tests, but here the treatment effect was also significant for figural originality on the Picture Completion sub-test and for verbal originality on the Product Improvement sub-test. The effect of classes within-treatments was significant for figural fluency on the Lines and Picture Completion sub-tests, for figural flexibility on the Lines sub-test, and for verbal originality on the Unusual Uses sub-test. The significant classes-within-treatments effect indicates differences among classes in the effectiveness of the program.

The adjusted means for experimental treatments for each of the dependent variables are presented in Table 2. The Newman-Keuls analyses revealed that the differences between control and treatment conditions, as measured by the Lines and Unusual Uses sub-tests, were significant for all creativity dimensions, with treatment means being greater than control means in all instances. The comparisons between the two experimental conditions, program with and without reinforcement, revealed that the differences were significant for figural fluency, flexibility and originality, and for verbal fluency. In all instances, including the two dependent variables for which the difference was not significant, the means for the non-reinforcement condition were greater than the means for the reinforcement condition.

Discussion

The results of this research confirm the effectiveness of creativity training in another culture, especially with

Table 2.—Adjusted Means for Creativity Test Scores

Experimental Condition	Figural (L = Lines, PC = Picture Completion)						Verbal (U = Unusual Uses, PI = Product Improvement)					
	Fluency		Flexibility		Originality		Fluency		Flexibility		Originality	
	L	PC	L	PC	L	PC	U	PI	U	PI	U	PI
Reinforced	16.31	9.40	11.20	8.15	18.04	9.88	17.45	13.51	6.05	4.78	5.37	8.38
Not reinforced	17.94	9.46	12.24	8.14	19.84	9.77	20.21	12.28	6.10	4.85	5.64	7.30
Control	14.45	9.42	10.26	8.22	13.86	9.13	14.52	12.79	4.03	4.73	2.32	7.81

children in a relatively underdeveloped educational environment. Memorization, respect for authority, and quite rigid teacher control are stressed in Brazilian schools, and classes are large. Such conditions might be expected to make children impervious to the effects of a creativity training program. Furthermore, the teachers were quite unfamiliar with the basic approach of the Purdue Creative Thinking Program which encourages children to develop their own ideas, to feel free to explore new concepts, and to be original. In spite of the handicaps the children and teachers faced, the gains in creative thinking ability were substantial. Whether the differential gains among classes was a teacher effect is not clear, but this possibility exists.

The results also indicate that the children who received reinforcement on their written productions in the form of encouraging comments performed less well than children who received no comments. It seems at first glance that reinforcers, in the form of brief comments, should augment learning, as Page (13) found true with elementary school children. However, reinforcers in the form of brief comments written on papers may evoke a sense of being evaluated and attendant fears or anxiety. As suggested by Osborn (12), Wallach and Kogan (20), and by Feldhusen and Hobson (9), freedom from concern about being judged or evaluated may be the best environment to foster not only immediate creative production but also longer-range learning of creative thinking. The latter proposition seems to be confirmed by the results of this research.

The Torrance Tests have been criticized as not having well-established validity, and particularly as not having a clearly established relationship with real life creative production (2). However, Treffinger (19) reviewed the research related to creative thinking and concluded that there is substantial predictive and concurrent criterion-referenced validity for creativity measures.

Thus, it seems likely that creativity training can be profitable for children in underdeveloped countries, that the Purdue Creativity Training Program may be particularly effective, and that the Torrance Tests of Creative Thinking can be used as one set of criterion measures. It is hoped that further research will be conducted to establish the generalizability and applicability of these results in children's lives outside of school and in later stages of their lives.

REFERENCES

1. Covington, Martin V.; Crutchfield, Richard S.; and Davies, Lillian, *The Productive Thinking Program,* Merrill, Columbus, Ohio, 1972.
2. Crockenberg, Susan B., "Creativity Tests: A Boon or Boondoggle for Education?," *Review of Educational Research,* 42:27-45, 1972.
3. Davis, Gary A., *Psychology of Problem Solving,* Basic Books, New York, 1973.
4. Demos, George D.; Gowan, John C.; and Torrance, E. Paul, *Creativity: Its Educational Implications,* Wiley, New York, 1967.
5. Feldhusen, John F.; Bahlke, Susan J.; and Treffinger, Donald J., "Teaching Creative Thinking," *Elementary School Journal,* 70:48-53, 1969.
6. Feldhusen, John F.; Treffinger, Donald J.; and Bahlke, Susan J., "Developing Creative Thinking: The Purdue Creative Thinking Program," *Journal of Creative Behavior,* 4:85-90, 1970.
7. Feldhusen, John F.; Treffinger, Donald J.; and Bahlke, Susan J., *Global and Componential Evaluation of Creativity Instructional Materials,* Creative Education Foundation, Buffalo, N.Y., 1970.
8. Feldhusen, John F.; Speedie, Stuart M.; and Treffinger, Donald J., "The Purdue Creative Thinking Program: Research and Evaluation," *NSPI Journal,* 10:5-9, 1971.
9. Feldhusen, John F.; and Hobson, Susan K., "Freedom and Play: Catalysts for Creativity," *Elementary School Journal,* 73:149-155, 1972.
10. Guilford, J. Paul, *The Nature of Human Intelligence,* McGraw Hill, New York, 1967.
11. Houtz, John C.; and Feldhusen, John F., "The Behavior Modification of Fourth Graders' Problem Solving Ability by Use of the Premack Principle and Special Instructional Material," Technical Report, OEG-52-9094, U.S. Office of Education, 1974, 226 pp.

12. Osborn, Lyle E., *Applied Imagination,* Scribner's, New York, 1963.
13. Page, Ellis, B., "Teacher Comments and Student Performance," *Journal of Educational Psychology,* 49:173-181, 1958.
14. Shively, Joe E.; Treffinger, Donald J.; and Feldhusen, John F., "Developing Creativity and Related Attitudes," *Journal of Experimental Education,* 41:63-69, 1972.
15. Speedie, Stuart M.; Treffinger, Donald J.; and Feldhusen, John F., "Evaluation of Components of the Purdue Creative Thinking Program: A Longitudinal Study," *Psychological Reports,* 29:395-398, 1971.
16. Torrance, E. Paul, *Torrance Tests of Creative Thinking,* Personnel Press, Princeton, N.J., 1966.
17. Torrance, E. Paul, *Rewarding Creative Behavior: Experiments in Classroom Creativity,* Prentice-Hall, Englewood Cliffs, N.J., 1967.
18. Torrance, E. Paul; and Myers, Robert E., *Creative Learning and Teaching,* Dodd, Mead, New York, 1970.
19. Treffinger, Donald J., "Assessment of Creative Problem Solving," Paper presented at the meeting of the American Psychological Association, Honolulu, 1972.
20. Wallach, Michael A.; and Kogan, Nathan, *Modes of Thinking in Young Children,* Holt, Rinehart and Winston, New York, 1965.

HEURISTICS FOR CLASSROOM DESIGN

CHARLES W. LAMB, JR.
Texas A. & M. University

ABSTRACT

Educators have long debated the question of optimal social design for the classroom. Particularly controversial have been the issues of leadership style and structural design of the classroom or learning environment. Perhaps the best answer to the question, "What is the 'optimal' social design that can be applied in the classroom?" is, "It all depends." The purpose of this paper is to present a set of heuristics which provides a systematic approach for developing social designs for classroom learning under various contingencies.

THE APPROPRIATE STARTING POINT for any approach to developing social designs for classroom learning is the educational objective. It is generally accepted that intellectual growth is a primary objective of education. Intellectual growth includes: (*a*) learning new information; (*b*) new applications and/or techniques; and (*c*) motivation.

The introduction of new information is generally thought of as descriptive or technical in nature. Introductory-level courses are usually designed to provide this form of intellectual growth.

Learning new applications or techniques often entails the utilization of principles or concepts previously learned. Examples of this form of intellectual growth include case studies, problem solving, and laboratory research.

Motivation is somewhat more abstract. Motivation may be viewed as stimulation of thought and ideas which generates interest for further intellectual growth in an area or field.

Any one or a combination of these forms of growth may be the objective of a specific academic course. The purpose of this paper is to present an approach which may be used to determine the most appropriate social design for achieving these forms of growth, given certain input variables.

Inputs

Two input variables are considered: (*a*) the belief system of the students; and (*b*) their previous educational backgrounds.

Several researchers have demonstrated that it is possible to categorize individuals based upon their belief system. The work of Rokeach (3), DiRenzo (2), and Stern, Stein, and Bloom (5) can be consolidated to characterize dogmatic and non-dogmatic students as follows:

Dogmatic	Non-Dogmatic
Authoritarian	Non-Authoritarian
Closed-minded	Open-minded
Rigid in opinions and beliefs	Flexible in opinions and beliefs
Low tolerance toward others	High tolerance toward others
Inconsiderate of others	Considerate of others
Conservative	Liberal
Depersonalized relationships with others	Highly personalized relationships with others
Inhibited	Outgoing
Average intelligence	Above average intelligence
Requires highly structured environment	Excels in loosely structured environment
Prefers lecture method of instruction	Prefers discussion method of instruction

Prepares for exams by memorizing main points	Prepares for exams by trying to understand concepts
Prefers objective tests	Prefers essay exams
Participates about average in class	Active participant in class discussions
Is chiefly a critic and evaluator	Is chiefly a contributor of ideas and concepts
Vocational interests in engineering, physical sciences, law, accounting, etc.	Vocational interests in social sciences, humanities, teaching, etc.

While it is not the objective of this paper to discuss methodologies for classifying classroom groups in terms of dogmatism, it is worthwhile to point out that this is indeed possible, and proven tests for this determination are available. Examples include the Rokeach Dogmatic Scale, the California F Scale, the Gough-Sanford Rigidity Scale, the Opinionation Scale, and the Ethnocentric Scale (4).

The second input variable, previous educational background, refers primarily, but not exclusively, to the level of attainment an individual has reached in a given area or field. More will be said about this variable later.

Social Designs

The social design of a classroom learning situation is primarily composed of two variables, the leadership style of the teacher and the structural design of the class.

Leadership style is generally viewed as a dichotomy. The labels vary somewhat but the dichotomous nature of styles is prevalent in the literature. Anderson, in a review of research concerning the effects of leadership styles, synthesized a large number of studies which referred to "authoritarian" versus "democratic" and "teacher-centered" versus "student-centered" styles (1). Although this construct is extremely oversimplified, it tends to be quite widely applied in analyzing leadership styles.

In considering structural designs it is important to note the various interaction roles of both faculty member and students. The structural design of the class can be one or a combination of the following:

1. *Lecture*–In this case the faculty member assumes the traditional role of teacher.
2. *Teacher-led discussion*–Here, at least conceptually, the faculty member assumes the role of a guide. Exchange of ideas is primarily of a multiple bilateral nature.
3. *Teacher as mentor*–Somewhere between teacher-led discussion and open discussion is a situation in which the faculty member assumes the role of a mentor. Conceptually this would compare favorably with an advanced undergraduate- or graduate-level seminar. Interaction is multilateral by design.
4. *Open discussion*–Here again interaction is multilateral but entirely among colleagues. The structure exhibits an obvious absence of a central authority figure.
5. *Independent study*–In this case the student is basically alone in his search for knowledge.

The descriptive material thus far presented can be easily summarized in terms of the model or flow diagram of Figure 1.

Figure 1.–A model for classroom design

The model is intended to indicate flow commencing with the given inputs and resulting in intellectual growth. Unfortunately, social design is often considered the equivalent of the "black box" concept. The conceptual framework of social designs is poorly understood by many academicians and therefore often applied incorrectly.

Instead of attempting to hypothesize the effect on intellectual growth of each possible combination of social designs given the possible input parameters, the remainder of this paper will be devoted to synthesizing the relevant literature in terms of the model, and to developing generalizations which may assist the teacher in formulating social design heuristics and policies for specific classroom situations.

The Student

The literature on individual characteristics strongly supports the proposition that a dogmatic individual, regardless of previous educational experiences, will achieve highest in a structured environment. This would include autocratic (or its semantic equivalent) leadership style and rigidly structured lectures. Whether functioning at an introductory or advanced level, within or outside of his area of specialization, the dogmatic individual excels when course content is explicit and/or technical in nature. Likewise, he dislikes and tends to do poorly in courses which he feels have little relevance to his specific career objectives. He views education as vocational preparation and is not likely to be interested in subjects outside of his primary field(s).

Assuming that a "class personality" does exist and that it is predominantly dogmatic, the optimal social design is apparent. Intellectual growth will best be faciliated by authoritarian leadership with the structural design emphasizing lectures in a factual, explicit manner.

At the other extreme of individual type is found the non-dogmatic group. Previous educational experiences and the intellectual growth goal of the specific course are

important factors to consider in determining the best social design to apply in this situation.

Previous Educational Background

There are generally two broad categories of previous educational experience. At one extreme are students who have had no formal exposure to a particular area of study. Students enrolled in introductory courses fall into this category. At the other extreme are students who have had educational experiences which directly support and provide background information for the particular course in which they are enrolled. Typical of this group is an advanced undergraduate or graduate student taking a course in his major or minor area.

Intellectual Growth Objective

Given that the student group is non-dogmatic, the intellectual growth objective(s) become important criteria in selecting the optimal social design.

In the case of the student with no previous exposure in an area, the intellectual growth objective is normally the introduction to new information. Most studies suggest that if new information is the purpose of a course, an "instructor-centered" leadership style is most effective. Also most appropriate in this situation is either a lecture or a combination of lecture and teacher-led discussion structure. Students in this group are being exposed to an entirely new environment; they are more at ease and learn better when their primary responsibility is to listen and absorb the new information in straight-forward, explicit terms. This also indicates their expectations in the course. They feel that neither they nor their colleagues have much to contribute to the group's learning process, and thus feel a strong need for a central leader-authority to guide them. Their discussions are normally of a multiple bilateral nature, with each student interacting only with the leader. Other structural designs tend to result in the same type of interaction.

Next, one can generalize using the other end of the educational experience spectrum and again look at the growth objective through the nature of the course. The typical course in this category will be advanced and its content less explicit. The course is normally an integral part of the student's curriculum. Therefore, both the students and the teacher have higher expectations. The intellectual growth objectives emphasize applications, techniques, and motivation. The most effective leadership style in this case will be "non-directive" or "student-centered." Applicable structural designs are flexible, from teacher-led discussion to open discussion to independent study.

The generalizations thus far presented are well documented in empirical studies. The realization that there is some discrepancy as to the optimal structural design in a given situation has provided the motivation for many of the research projects which formed the foundation for this investigation.

Summary and Conclusion

The objective of this paper has been to provide a systematic approach for developing social designs for classroom learning under various contingencies. The recommended procedure is as follows:

1. Determine inputs
2. Determine intellectual growth objective(s)
3. Determine optimal structural design
4. Determine appropriate leadership style

It would be a rare case to find the variables discussed in this paper perfectly uniform in a classroom situation. It is therefore not possible to prescribe absolute formulas for classroom success. It is, however, possible to provide a systematic approach for analyzing the objectives, parameters, and variables associated with intellectual growth.

The utility of this or any other conceptual approach can only be determined by empirical testing. If it provides the framework for a logical decision process in the development of social designs for classroom learning, it will have achieved the stated objective.

REFERENCES

1. Anderson, R.C., "Learning in Discussion: A Resume of the Authoritarian-Democratic Studies," *Harvard Educational Review,* 29:201-215, 1959.
2. DiRenzo, G.J., *Personality, Power, and Politics,* University of Notre Dame Press, Notre Dame, Ind., 1967.
3. Rokeach, M., "The Nature and Meaning of Dogmatism," *Psychological Review,* 61:194-204, 1954.
4. Rokeach, M., *The Open and Closed Mind,* Basic Books, New York, 1960.
5. Stern, G.G.; Stein, M.I.; and Bloom, B.S., *Methods in Personality Assessment,* Free Press, Glencoe, Ill., 1956.

ITEM-BY-ITEM FEEDBACK AND MULTIPLE CHOICE TEST PERFORMANCE[1, 2, 3]

R. STEPHEN FULMER
Bristol Regional Mental Health Center
Bristol, Tennessee

HARRY E. ROLLINGS
Marquette University

ABSTRACT

The reinforcing and/or informational effects of immediacy of feedback on test performance have been a topic of some debate for several years. In the present research, it was hypothesized that subjects receiving knowledge of the correct answer to each item, immediately after answering, would exhibit superior test performance when compared with control groups. Subjects were matched-paired on the basis of their first examination scores. Subsequently, one group was tested on a Modular EDEX Student Response System which provided immediate item-by-item feedback. Control groups took the examination in the traditional and modular manner with no immediate feedback. Scores of the immediate feedback group were significantly higher than those of the traditional groups.

THE EFFECTS ON PERFORMANCE of the delay of feedback in classroom examination situations has been the topic of considerable investigation. Some researchers have adopted a general position that feedback should come as soon as possible after the response due to a negative relationship between delay of feedback and effectiveness of learning and retention (1). However, several later studies take issue with this position. English and Kinzer (4) found a feedback delay of two days to be more effective than immediate feedback in tests of the retention of new material. More (6) found that delays of 2½ hours and one day increased the acquisition rate of new material over immediate feedback, and concluded that immediate feedback "may not only be ineffective, but may actually inhibit retention learning." Beeson (2), in support of the superiority of immediate feedback, found that under certain testing conditions groups receiving immediate feedback performed significantly better than those receiving delayed feedback.

In an attempt to resolve these conflicting findings, the present research was designed to study the effect of immediate item-by-item feedback versus no feedback on multiple choice test performance. The present investigators viewed feedback in the more affective terms of reinforcement, and controlled for several possible sources of confounding, such as item dependency, machine-novelty, and different room/different instructor effects. It was predicted that students receiving immediate feedback would score significantly higher on the examination than students receiving no feedback during testing.

Method

Apparatus

A Modular EDEX Student Response System was used which allowed the instructor to provide immediate oral feedback for up to 40 students simultaneously. Each student had a response unit with five buttons and could respond by pushing one of the buttons A–D or, if he desired to make no response, could push the blank button. These units were placed on standard classroom desks. The instructor's monitor had the capacity to: (*a*) give the percentage of students making each response; (*b*) indicate the response made by each student; and (*c*) record the total number of correct responses for each student. Groups using the above apparatus were termed *machine*, while those that did not were termed *traditional.*

Subjects

Eighty-four students from two Introductory Psychology classes plus thirty students from one Child Psychology class at Middle Tennessee State University served as *S*s. The Introductory students were divided into four groups (two from each class) of twenty-one *S*s each, and the Child Psychology class was divided into two groups of fifteen. The lack of another Introductory Psychology class necessitated using the Child Psychology class. All classes were taught by the same instructor on the same days.

Procedure

Test 1 was given by the traditional (mimeographed) method to all students from the Introductory classes, and consisted of sixty multiple choice (four alternative) items on material covered in the course. On the basis of Test 1 scores, the students were match-paired into four groups. One of the Introductory classes was subdivided into two groups: (*a*) the immediate feedback group which received item-by-item feedback during Test 2 and (*b*)the feedback control group which took Test 2 in the traditional manner The feedback control group took Test 2 at the same time of day as the immediate feedback group but in a different room proctored by a different professor. They received mimeographed copies of Test 2 and recorded their answers on IBM answer sheets. In another room, the immediate feedback group received the same mimeographed test, but they recorded their answers on the EDEX System and were given feedback following each response.

Group I *S*s from the Introductory class took Test 2 in the the traditional manner (using IBM answer sheets) in their usual classroom and were proctored by the regular instructor. At the same time, Group II from this class took Test 2 in the traditional manner, but in a new classroom with a different instructor. Thus, on Test 2 there was one experimental group (feedback) and three control groups (feedback control, same-instructor/same-room control, and different-room/different-instructor control). In order to investigate the possibility of a machine-novelty effect, the Child Psychology class was divided into two groups: a machine no-feedback group and a no-feedback control group.

Control groups for different-room and different-teacher effects are suggested by previous research (3, 5), which indicates that test performance and learning differences may arise between groups due to contextual changes in the learning environment, such as the room where original and interpolated learning took place.

Results and Discussion

The results of the machine-novelty comparison indicate that while there was a slight difference in favor of the machine no-feedback group over its traditional test administration method control, the difference was not statistically significant ($t < 1.0$). This indicates that any differences found in the machine no-feedback group versus the traditional test administration method would not be a function of the novelty of the machine itself.

Table 1.–Mean Correct Responses for Introductory Psychology Classes Tested

Group	Old Teacher Old Room	New Teacher New Room	Machine Feedback	Machine No-Feedback
Standard testing	31.37 (8.66)*	31.00 (8.61)	31.95 (8.52)	31.68 (9.07)
Treatment testing	33.26 (8.64)	33.26 (8.54)	38.55 (6.79)	34.18 (10.04)
N	21	21	21	21

*Standard deviations in parentheses

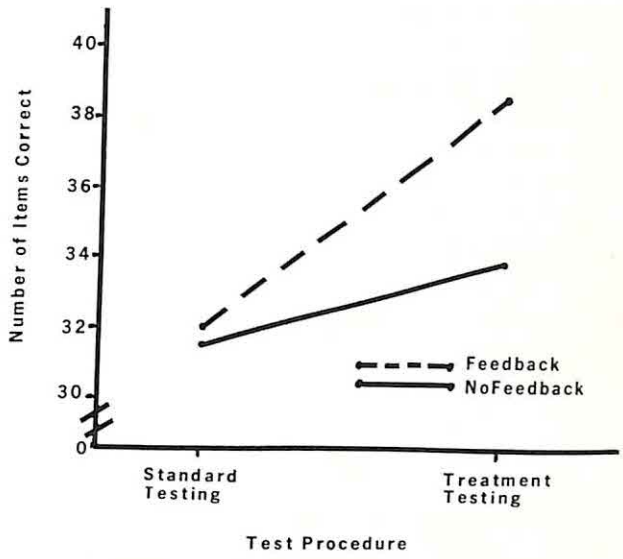


Figure 1.–Comparison of Test 1 standard testing with Test 2 feedback versus no-feedback testing

One factor that could have produced a difference on Test 2 was that students taking a test in a new room proctored by a new professor might have behaved differently than if they had been in their regular classroom with their regular professor. However, there was no differential effect due to teacher-room differences on multiple choice test performance ($t < 1.0$).

With the effects of both machine and teacher-room difference accounted for, it seems reasonable to conclude that any difference on Test 2 for the feedback versus the three control groups might be due to the effects of immediacy of feedback. Figure 1 illustrates the effect of feedback versus the no-feedback control on test performance. The mean difference of 4.5 more correct items in the feedback group over the three control groups was reliable, $F(3, 60) = 4.01; p < .025$. Table 1 presents the means and standard deviations for the 84 *S*s in the Introductory Psychology course. Using the Newman-Keuls method for testing individual mean differences, no differences were found among the control groups.

Students receiving rapid item-by-item feedback on multiple choice examination items performed better than those receiving no feedback by approximately 7.5%. One factor that may have been responsible for the higher performance of the feedback group was test dependency (i. e., knowledge of the correct answer to one item could

have given away all or part of the answer to a future item). In order to circumvent this potential source of confounding, five different instructors were asked to independently analyze Test 2 and point out any items that were, in their opinion, dependent. There were sixteen items which at least three of the five instructors judged as possibly having some degree of dependence. When these items were excluded from data analysis, the previous results were again obtained. Thus, one can reasonably assume that the previous difference was probably due to feedback and not item dependence.

NOTES

1. Portions of the data were presented at the meeting of the Southeastern Psychological Association, New Orleans, May 1973.
2. This research was supported in part by the Marquette University Committee on Research.
3. Reprint requests should be sent to R. Stephen Fulmer, Bristol Regional Mental Health Center, 26 Midway Street, Bristol, Tennessee 37620.

REFERENCES

1. Ammons, R. B., "Effects of Knowledge of Performance: A Survey and Tentative Theoretical Formulation," *Journal of General Psychology*, 54: 279-299, 1956.
2. Beeson, R. O., "Immediate Knowledge of Results and Test Performance," unpublished doctoral dissertation, University of Arkansas, 1970.
3. Bilodeau, I. McD.; and Schlosberg, H., "Similarity in Stimulating Conditions as a Variable in Retroactive Inhibition," *Journal of Experimental Psychology*, 41:199-204, 1951.
4. English, R. A.; and Kinzer, J. R., "The Effects of Immediate and Delayed Feedback on Retention of Subject Matter," *Psychology in the Schools*, 3: 143-147, 1966.
5. Greenspoon, J.; and Ranyard, R., "Stimulus Conditions and Retroactive Inhibition," *Journal of Experimental Psychology*, 53: 55-59, 1957.
6. More, A. J., "Delay of Feedback and the Acquisition and Retention of Verbal Materials in the Classroom," *Journal of Educational Psychology*, 60: 339-342, 1969.

THE EFFECT OF HUMAN RELATIONS TRAINING ON DOGMATIC ATTITUDES OF EDUCATIONAL ADMINISTRATION STUDENTS

JOHN MORACCO
ABDUL-GHANI BUSHWAR
American University of Beirut
Beirut, Lebanon

ABSTRACT

The effects of human relations seminars on dogmatism scores of educational administration students were investigated. Two groups of graduate students coming largely from the Middle and Far East were used in this study. The experimental group was given a series of human relations workshops over a period of four weeks. The results provided evidence that human relations training can reduce dogmatism as measured by the Rokeach Dogmatism Scale. Also, the seminars provided an opportunity not usually found in the traditional academic program for graduate students to interact with their peers and the faculty.

TRAINING STUDENTS TO BE competent teachers and school administrators is one of the most serious tasks that educators must undertake. Accordingly, efforts are continuously being made by educators to devise programs that will have a positive influence on their students' atti-

To this end studies carried out in the past decade have attempted to investigate the personality factors and dimensions of teachers and school administrators. Research on dogmatism has demonstrated that open-minded instructional leaders exhibit behaviors which are consistent with the goals of providing a democratic atmosphere conducive to learning by inquiry (1). Further, other researchers have presented evidence to suggest that open-minded individuals have the requisite attitudes that characterize effective teachers and administrators (4-6).

Recently, human relations training has been promoted as a promising method of reducing dogmatic attitudes (2,8,9). In these studies the reduction of dogmatic attitudes has been accomplished through group experiences. Exposure to human relations training can help in promoting desirable attitudes and is consistent with some theories of educational administration (3).

The purpose of this study was to ascertain whether dogmatism of prospective school administrators coming largely from the Middle and Far East can be modified through exposure to a group experience in human relations. Additionally, the study was designed for a relatively short period of time to test whether changes in attitudes can be accomplished quickly.

Method

Subjects

The *Ss* used in this study were three female and thirteen male graduate students registered for the Spring 1974 semester in a program leading to a master's degree in Educational Administration at the American University of Beirut. The *Ss* were randomly assigned to either the experimental or the control group. The experimental group consisted of two females and six males, with an average age of 34 years; the control group consisted of one female and seven males, and the average age was 35 years. For both groups, most of the *Ss* were married and all subjects had previous teaching experience.

Instrument

The measuring instrument used in this study was an adaptation of Rokeach's Dogmatism Scale (Form E). Form E was specifically designed and validated to assess the degree of open- and closed-mindedness in an individual (5). Rokeach (7) states that the instrument can be used to measure general authoritarianism and intolerance.

Form E was adapted for use for the present study by conducting an item analysis on a pilot sample of 75 *Ss* in the Education Department. Of the original 40 items, 11 items did not discriminate for the sample and were subsequently dropped from the scale. This procedure yielded an adapted scale of 29 items with a odd-even reliability coefficient of .85. This was considered sufficient for the purpose of the study.

The *Ss* were required to indicate their feelings about each item on a 6-point forced-choice scale. The choices ranged from "I agree very much" (+ 3) to "I disagree very much" (- 3). A constant of 4 was added to each item score to eliminate negative numbers. A higher score on the Dogmatism Scale indicates a stronger dogmatic attitude.

Procedure

The *Ss* in the study were taking basically the same course work as they were all approximately at the same stage in their graduate work. It was announced in class that some students would be chosen to participate in small group discussions outside of class. The purpose was given as an opportunity to become acquainted with class members in another setting. No credit was given for participation, and attendance after the experimental group was chosen was voluntary. The experimental group attended all other regularly scheduled classes in addition to the human relations sessions. The control group attended only the regularly scheduled classes.

The two groups, experimental and control, of eight members each responded to the revised Dogmatism Scale as a pre-test. The means were tested and the difference was found not to be significant, as seen from Table 1. The experimental group was exposed to eight 1½-hour human relations groups meeting over a period of four weeks. The leader of the group, a faculty member not previously exposed to the *Ss*, had a Ph.D. in counseling and was a trained group leader.

The human relations group was basically unstructured, had no fixed agenda, and stressed the expression of feelings and ideas experienced within the group and/or outside of it. The topics of discussion ranged from very personal problems to academic and on-the-job problems. An effort was made by the leader to relate the discussion as much as possible to the school administrator's role. The topics and the feelings that they generated were held confidential by the experimental group, thus helping to develop trust within the group as well as to keep the interaction between the control and the experimental groups at a minimum.

At the end of four weeks a post-test was administered to both the experimental and control groups. Tables 1 and 2 show the results of the correlated and independent *t*-tests that were computed.

Results

From Table 1 it can be seen that the experimental group and the control group did not significantly differ on their pre-test scores. From Table 2 it can be seen that the experimental group's mean post-test score was significantly less than its pre-test mean score, whereas the control group's mean post-test score was greater (though not significant) than its pre-test score. Further, it can be seen in Table 1 that the experimental group's post-test mean score was significantly lower than the control group's post-test mean score.

Table 1.–Summary of Independent *t*-Tests of Control and Experimental Groups on Pre- and Post-Tests

Test	Experimental Group (N=8)	Control Group (N=8)	Independent *t*-value (df=14)
Pre-test	$\bar{X}_1$ = 123.13	$\bar{X}_1$ = 119.13	1.09*
	SD_1 = 7.61	SD_1 = 7.00	
Post-test	$\bar{X}_2$ = 114.63	$\bar{X}_2$ = 120.88	
	SD_1 = 5.48	SD_2 = 5.74	2.22**

*Significant at the .01 level
**Significant at the .05 level

Table 2.–Summary of Correlated *t*-Tests of Pre- and Post-Test Means for the Experimental and Control Groups

Group	Pre-test	Post-test	Correlated *t*-value (df=7)
Experimental (N=8)	$\bar{X}_1$ = 123.13	$\bar{X}_2$ = 114.63	
	SD_1 = 7.61	SD_2 = 5.48	4.49*
Control (N=8)	$\bar{X}_2$ = 119.13	$\bar{X}_2$ = 120.88	
	SD_1 = 7.00	SD_2 = 5.74	1.45**

*Significant at the .01 level
**Significant at the .05 level

An open-ended questionnaire was given to the experimental group as a means to obtain some feedback on the participants' reactions to the human relations sessions. A summary of the responses follows.

Discussion

The results of this study provided evidence that dogmatic attitudes can be changed by human relations training over a relatively short period of time. Moreover, by summarizing the responses of the open-ended questionnaire it was found that group experiences should become a formal part of the academic program in educational administration. The participants felt that the human relations group provided a vehicle for learning that was not present in their regular classroom. Also, the experimental group felt that they were able to get to know their classmates in a more intimate way through the group experience, and that the experience was meaningful enough that they recommended human relations experiences as a regular part of the graduate program. It gave them an opportunity to develop contacts with their peers in a manner not normally found in traditional programs. The experience also gave the group the opportunity to interact with the group leader, a member of the faculty, whose concern about their welfare was highly appreciated by the students.

Whether the changed attitudes of the experimental group are permanent and whether these attitudes are consistent with their actual behavior as school administrators are important considerations for further investigation. Research should be done with larger groups, and it should have a component which assesses actual behaviors in schools.

Recent evidence has been accumulating to show that group experiences should become an integral part of the training of school administrators. Savage (8) has been a strong proponent of this approach. This new emphasis is certain to become even stronger in the future than it is now.

REFERENCES

1. Ager, M., "Dogmatism and the Verbal Behavior of Student Teachers," *The Journal of Teacher Education*, 21:179-183, 1970.
2. Bunker, D., "The Effects of Laboratory Education upon Individual Behavior," in Schein and Bennis (eds.), *Personal and Organizational Change through Group Methods*, Wiley, New York, 1965.
3. Coladarci, A.P.; and Cetzels, J.W., *The Use of Theory in Educational Administration*, Stanford University Press, Palo Alto, Calif., 1955.
4. Gregg, D.B., "Keys to Effective Behavior," *The Journal of Teacher Education*, 22:464-468, 1971.
5. Johnson, J.S., "Dogmatism: A Variable in the Prediction of Student Teaching Performance," *Contemporary Education*, 41:14-18, 1969.
6. Musella, D., "Open/Closed-Mindedness as Related to the Rating of Teachers by Elementary School Principals," *The Journal of Experimental Education*, 35:75-79, 1967.
7. Rokeach, M., *The Open and Closed Mind: Investigations into the Nature of Beliefs Systems and Personality Systems*, Basic Books, New York, 1960.
8. Savage, W.W., *Interpersonal and Group Relations in Educational Administration*, Scott, Foresman, Glenview, Ill., 1968.
9. Wilson, J.E.; Morton, R.B.; and Mullen, P.P., "The Trend in Laboratory Education for Managers: Organization Training or Sensitivity?," *Training and Development Journal*, 26:18-25, 1972.

THE TWO EDITIONS OF SOME INTRODUCTORY PSYCHOLOGY TEXTBOOKS

M.Y. QUERESHI
MICHAEL R. ZULLI
Marquette University

ABSTRACT

A comparative content analysis of the index terms employed by two different editions of twelve introductory psychology textbooks, utilizing principal components analysis with varimax rotation, revealed a definite trend toward more uniformity among the textbooks with respect to the terms employed in the later versions than those used in the earlier editions of these texts. The relative prominence of various areas of psychology, however, remained about the same in both editions of the selected textbooks. The Spearman rho of .72 ($p < .01$) between the number of prominent terms employed in the two editions indicated that the revision did not result in any substantial change in the relative status of the texts in regard to the thoroughness of their indices.

THE PAST TWO DECADES have witnessed the extensive growth of psychology both as a science and as a profession. Similarly, the popularity of psychology courses taught at the college level has increased substantially. In the undergraduate curriculum, the introductory psychology course is probably among the most commonly offered courses since it is offered by 92% or more of the universities, liberal arts colleges, and junior colleges (14:64). To meet this demand, an increased number of textbooks have been written for the introductory course, and these textbooks have been revised usually with greater frequency than textbooks for other psychology courses.

The contents of these textbooks are subjectively evaluated by experts and these reviews are published in *Contemporary Psychology* and other appropriate sources (1). However, few attempts have been made to analyze systematically the contents (e.g., terms employed in the texts as reflected in the subject indices) in order to determine their similarity or relative comprehensiveness. In addition, one should inquire whether revisions of these introductory texts represent any improvement, at least with respect to the convergence of the prominent terms that constitute the core of the introductory psychological literature.

A previous study (23) dealt with the first of the two aforementioned questions. The present study was chiefly concerned with the second problem, namely, systematically comparing the contents of the two different editions of some introductory psychology textbooks.

Method

Sample

Of the 25 introductory psychology textbooks that the authors were able to obtain from the faculty, as well as from the publishers, in the fall of 1972, twelve had been revised between 1968 and 1972. Depending on availability, one of the past editions of each of these books was then borrowed from either local libraries or faculty members. The terms listed in the subject indices of the two different editions of these twelve books (2-13, 15-22, 24, 25, 27, 28) were treated as two separate content domains for the purpose of this analysis.

Procedure

The following procedure was followed for each edition: First, each text was assigned a two-digit identification number. Second, lists were prepared of all terms in the main headings of the indices, and the identification numbers of the textbooks in which a term appeared were noted beside each term. Third, after alphabetizing the terms, each term was assigned a four-digit identification number and was punched on an IBM card. Each card representing a distinct term contained the identification number of the term, the scores of 1 or 0 in the subsequent columns (depending on whether that term was used in any of the twelve books of a particular edition), the total score for the term (number of times the term was used), and,

Table 1.–Correlations among Twelve Books in the Old (*above diagonal*) and New (*below diagonal*) Editions.

	1	2	3	4	5	6	7	8	9	10	11	12
1	1.00	.30	.30	.15	.25	.15	.20	.30	.10	.15	.15	.20
2	.30	1.00	.15	.30	.35	.25	.25	.40	.25	.35	.20	.20
3	.20	.15	1.00	.10	.25	.20	.10	.10	.10	.10	.20	.10
4	.20	.25	.20	1.00	.30	.15	.15	.20	.20	.20	.20	.15
5	.40	.30	.15	.20	1.00	.40	.35	.35	.30	.40	.35	.30
6	.25	.30	.15	.15	.30	1.00	.30	.25	.10	.10	.20	.10
7	.30	.25	.25	.25	.25	.15	1.00	.30	.20	.05	.20	.20
8	.25	.20	.20	.15	.20	.15	.35	1.00	.15	.20	.15	.15
9	.30	.15	.10	.15	.25	.15	.15	.15	1.00	.35	.20	.20
10	.15	.25	.20	.25	.10	.15	.20	.15	.10	1.00	.20	.20
11	.25	.20	.20	.20	.20	.15	.30	.20	.25	.20	1.00	.20
12	.20	.20	.20	.30	.20	.20	.30	.20	.15	.25	.20	1.00

finally, the term itself. Thus, the total number of terms was determined (5386 in the old and 6159 in the new edition), as well as their relative frequency for each edition separately.

Since many of the terms in each edition were used only once in a set of twelve books, it was decided to base the analysis only on those terms which were used two or more times in the texts in a particular edition. Thus, there were 1617 terms which were used twice or more in the old edition and 1895 terms used twice or more in the new edition of the same textbooks. The first data matrix (old edition), therefore, was of 1617 × 12 and the second, of 1895 × 12 size. Because of the dichotomous nature of the scores and the skewed shape of the distributions, tetrachoric correlations were computed to represent correlations among the books for each edition separately. Table 1 presents the correlations for the old edition (above diagonal) and new edition (below diagonal) textbooks. The two 12 × 12 correlation matrices (one for the old and the other for the new edition) were analyzed by means of the principal components method, using unities in the diagonal. Factors whose latent roots were not less than .80 were retained and rotated by the varimax routine, resulting in six principal components for each edition.[1] A second-order factor analysis, using the intercorrelation of six factors obtained through promax rotation (a method which yields oblique simple structure) was also conducted for the two editions separately. The methods of analysis and rotation were the same as before, but the retention of a factor depended on its having a latent root of 1.00 or larger.

The number of most prominent terms (used in six or more of the books in each sample) was also determined in order to compare the relative thoroughness of the two editions. Spearman's rank order coefficient was then computed between the number of prominent terms used in the old and those used in the new edition in order to determine the stability of the relative prominence of the books from one edition to another.

Finally, to determine on a rather broad scale whether the subject matter of introductory psychology changed qualitatively over a period of about ten years (the average time lag between the two selected editions), four broad areas were somewhat arbitrarily designated to cover the spectrum of psychological material: (*a*) psychometrics and statistics; (*b*) learning and physiological; (*c*) personality and psychopathology; and (*d*) social and developmental. By classifying the total number of most prominent terms from each sample (edition) into these four areas and then comparing the results of old and new editions, the authors hoped to secure a reasonably realistic impression of the change in emphasis on various subfields of psychology as reflected in the two different editions of these twelve books.

Results and Discussion

Factor Analysis of Textbooks

In accordance with the criterion of retention mentioned previously, six factors were retained and rotated for each edition separately. The factor loadings after rotation of the

twelve books in the old edition are presented in Table 2, while Table 3 presents analogous results for the new edition. The factors as a whole account for 69.7% of the total variance for the old edition (Table 2) and 67.4% of the total variance for the new edition (Table 3). The percentage of variance accounted for by various factors individually indicates that no single factor holds any dominant position in either analysis (the percentage of variance ranges between 9.2 and 13.7 in Table 2 and between 8.3 and 14.5 in Table 3). All of the factors in the first analysis (old edition in Table 2) can be justifiably designated as group factors since they have substantial loadings (± .25 or larger) on at least two but not more than five of the twelve textbooks. In the second analysis (Table 3), five are group factors, but one (Factor G) is a specific factor since it has substantial loading only on one book, Hilgard's 1957 edition (6).

To determine the similarity between the two sets of factors (Table 2 versus Table 3), Tucker's (26) coefficients of congruence were computed between the corresponding pairs (e.g., Factor A in Table 2 versus Factor A in Table 3, etc.). These coefficients for Factors A, B, C, D, E, and F were .69, .63, .84, .72, .69, and .44, respectively. Thus, there seems to be some similarity between the factors, but

Table 2.–Factor Loadings after Varimax Rotation of the Twelve Books (Old Edition)

Book	Factors						
	A	B	C	D	E	F	h^2
Edwards (1968)	-.04	-.04	.53	.61	.28	.01	.73
Hebb (1958)	.15	.35	.06	.63	.00	.26	.60
Hilgard (1957)	.12	.06	.88	.04	-.01	.03	.80
Kendler (1963)	.03	.08	-.02	.23	.03	.91	.89
Kimble (1956)	.51	.44	.22	.19	.16	.23	.62
Krech & Crutchfield (1958)	.78	.06	.19	.12	-.17	.07	.70
Lindgren & Byrne (1961)	.68	-.02	-.10	.26	.36	-.01	.66
McKeachie & Doyle (1966)	.32	.11	-.03	.72	.04	.06	.64
Morgan (1956)	.13	.73	-.03	.02	.18	.04	.59
Munn (1956)	-.05	.82	.10	.21	-.02	.10	.74
Ruch (1958)	.35	.23	.32	-.22	.26	.43	.57
Whittaker (1965)	.05	.19	.05	.09	.86	.09	.80
Percent of variance	13.5	13.7	10.6	13.0	9.2	9.7	69.7

Table 3.–Factor Loadings after Varimax Rotation of the Twelve Books (New Edition)

Book	Factors						
	A	B	C	D	E	F	h^2
Edwards (1972)	-.52	.41	.08	.29	.10	-.06	.54
Hebb (1970)	.59	.02	-.13	.20	.09	.47	.64
Hilgard, Atkinson, & Atkinson (1971)	.13	.07	.94	.14	.13	.09	.95
Kendler (1968)	.11	.12	.04	.03	.76	.18	.64
Kimble & Farmezy (1968)	.66	.26	.01	.17	.19	-.17	.59
Krech, Crutchfield, & Livson (1969)	.76	.04	.16	-.05	.06	.16	.64
Lindgren, Byrne, & Petrinovich (1968)	.10	.14	.12	.70	.29	.09	.63
McKeachie & Doyle (1971)	.14	.04	.05	.83	.02	.05	.72
Morgan & King (1971)	.17	.83	-.03	-.03	.12	-.06	.74
Munn, Fernald, & Fernald (1972)	.05	.08	.10	.06	.22	.82	.74
Ruch & Zimbardo (1971)	-.01	.62	.16	.27	.02	.37	.62
Whittaker (1970)	.12	.04	.08	.20	.76	.06	.64
Percent of variance	14.5	11.3	8.3	12.1	11.5	9.5	67.4

it is generally quite negligible except in the case of Factor C, where the coefficient of congruence is not interpreted as a correlation coefficient and usually has to be higher than .80 to indicate even a semblance of similarity between two factors. If the indices of these books, in the two selected editions, are fair indicators of the corresponding terminological contents, then the two editions would not be given a high rating on factorial congruence. However, it does not mean that the terms used in the index of the old edition of a book are different from those in the new edition of the same book, but it does indicate lack of stability between the two editions in the content areas represented by certain subgroups of texts.

The second-order analysis, based on six factors obtained through promax (oblique simple structure) solution and employing the principal components method with varimax rotation, resulted in two second-order factors for the data of the old edition and only one second-order factor for the new edition—the criterion for retention of a factor in both cases of the second-order analysis was having a latent root of 1.00 or larger. Table 4 embodies Factors 1 and 2 based on the old edition and one factor based on the texts in the new edition. While Factors 1 and 2 together account for 46.6% of the variance, the single factor emerging in the data of the new edition accounts for 39.5% of the variance. Thus, it seems that the new edition, in contradistinction to the old, evidences a greater degree of convergence and uniformity in the contents of the indices of the selected books than does the old edition. The books in the old edition do not seem to have as much common material among themselves as do their counterparts in the revised version. Over the years, therefore, authors have tended to employ a greater proportion of common terms in the indices of introductory texts.

Comparison of the Degree of Prominence

The degree of prominence of a term, in either edition, was defined as the inclusion of a term in the indices of six or more books. The twelve books in the old edition ranged from a low of 115 terms, used by Hebb (4), to a high of 259 terms, used by both Morgan (19) and Whittaker (27). The total number of prominent terms in the old edition was 305, compared with 368 in the new edition. The mean number of prominent terms in the old edition was 200.3, with a standard deviation of 51.5. The range for prominent terms in the new edition was from 150 by Hebb (5) to 295 by Hilgard, Atkinson, and Atkinson (7), which text included 255 prominent terms in its 1957 edition. The average increase in the number of terms in the new edition was 48.2, indicating that the later edition of these textbooks was much more comprehensive than the earlier one. Spearman's rho between the numbers of prominent terms used in the two editions of these twelve books turned out to be .72 ($p < .01$), indicating a substantial degree of stability in the relative prominence of the indices across two editions.

Table 4.—Results of the Second-Order Factor Analysis for the Old and New Editions

Variables	Factors ((Old Edition) 1	2	Factor (New Edition) 1
A	.62	.07	.77
B	.67	-.05	.65
C	.45	-.51	.66
D	.30	.83	.69
E	.48	.16	.48
F	.66	-.15	.46
Percent of Variance	29.9	16.7	39.5

Emphasis on Various Areas of Psychology

Classification by two research assistants of the prominent terms (305 in the old and 368 in the new edition) into four broad areas resulted in the following percentage-wise distribution: (*a*) psychometrics and statistics, 14% in both the old and new editions; (*b*) learning and physiological, 48% in the old and 52% in the new; (c) personality and psychopathology, 23% in the old and 19% in the new; and (*d*) social and developmental, 15% of the total terms in both editions. Thus, the relative emphasis given to the aforementioned areas was about the same in both editions, although the total number of prominent terms was much larger in the new edition of the books. In spite of the subjective character of the foregoing procedure, the results seem to conform to those of the first-order principal components analysis in which the same number of factors emerged in the data of both editions.

Evaluation and Conclusion

It is recognized that the present findings about the contents of these textbooks are valid to the extent to which the indices, in either edition, accurately represent the contents and organization of the twelve selected books. Some of the authors apparently devoted considerable care to the preparation of indices, while others did not. Also, the time gap between the two editions was not the same for all texts, especially since in some cases the authors were unable to secure the immediate past edition and in other cases the texts had gone through no more than two revisions. In general, these conclusions seem to make sense in the light of what is generally known about the quality of these texts and their revisions (1).

NOTE

1. Although a number of authorities on factor analysis recommend a latent root of 1.00 or larger for retaining a factor

for rotation, .80 was selected as the cutoff point for first-order analysis in this study because it permitted the accounting of about 70% of the variance by the components for the data of each edition. Had 1.00 been used as the cutoff instead of .80, the authors would have ended up with about 40% of the variance and with three, instead of six, components extracted in each case. On the other hand, changing the .80 criterion to a lower figure would have resulted in the retention of a number of additional components whose variance contributions were minimal.

REFERENCES

1. American Psychological Association, *Psychology Teacher's Resource Book,* Washington, D.C., 1973.
2. Edwards, D.C., *General Psychology,* Macmillan, New York, 1968.
3. Edwards, D.C., *General Psychology* (2nd ed.), Macmillan, New York, 1972.
4. Hebb, D.O., *A Textbook of Psychology,* Saunders, Philadelphia, 1958.
5. Hebb, D.O., *Textbook of Psychology* (2nd ed.), Saunders, Philadelphia, 1970.
6. Hilgard, E.R., *Introduction to Psychology* (2nd ed.), Harcourt, Brace, New York, 1957.
7. Hilgard, E.R.; Atkinson, R.C.; and Atkinson, R.L., *Introduction to Psychology* (5th ed.), Harcourt Brace Jovanovich, New York, 1971.
8. Kendler, H.K., *Psychology,* Appleton Century Crofts, New York, 1963.
9. Kendler, H.K., *Basic Psychology* (2nd ed.), Appleton Century Crofts, New York, 1968.
10. Kimble, G.A., *Principles of General Psychology,* Ronald Press, New York, 1956.
11. Kimble, G.A.; and Garmezy, N., *Principles of General Psychology* (3rd ed.), Ronald Press, New York, 1968.
12. Krech, D.; and Crutchfield, R.S., *Elements of Psychology,* Knopf, New York, 1958.
13. Krech, D.; Crutchfield, R.S.; and Livson, N., *Elements of Psychology* (2nd ed.), Knopf, New York, 1969.
14. Kulik, J.A., *Undergraduate Education in Psychology,* American Psychological Association, Washington, D.C., 1973.
15. Lindgren, H.C.; and Byrne, D., *Psychology: An Introduction to the Study of Human Behavior,* Wiley, New York, 1961.
16. Lindgren, H.C.; Byrne, D.; and Petrinovich, L., *Psychology: An Introduction to the Behavioral Science* (2nd ed.), Wiley, New York, 1968.
17. McKeachie, W.J.; and Doyle, C.L., *Psychology,* Addison-Wesley, Reading, Mass., 1966.
18. McKeachie, W.J.; and Doyle, C.L., *Psychology* (2nd ed.), Addison-Wesley, Reading, Mass., 1971.
19. Morgan, C.T., *Introduction to Psychology* (2nd ed.), McGraw-Hill, New York, 1956.
20. Morgan, C.T.; and King, R.A., *Introduction to Psychology* (4th ed.), McGraw-Hill, New York, 1971.
21. Munn, N.L., *Psychology* (3rd ed.), Houghton Mifflin, Boston, 1956.
22. Munn, N.L.; Fernald, Jr., L.D.; and Fernald, P.S., *Introduction to Psychology* (5th ed.), Houghton Mifflin, Boston, 1972.
23. Quereshi, M.Y.; and Zulli, M.R., "A Content Analysis of Introductory Psychology Textbooks," *Teaching of Psychology,* 1975, in press.
24. Ruch, F.L., *Psychology and Life* (5th ed.), Scott, Foresman, Chicago, 1958.
25. Ruch, F.L.; and Zimbardo, P.G., *Psychology and Life* (8th ed.), Scott, Foresman, Glenview, Ill., 1971.
26. Tucker, L.R., *A method for Synthesis of Factor Analysis Studies,* (Personnel Research Section Report No. 984), Department of the Army, Washington, D.C.. 1951.
27. Whittaker, J.O., *Introduction to Psychology,* Saunders, Philadelphia, 1965.
28. Whittaker, J.O., *Introduction to Psychology* (2nd ed.), Saunders, Philadelphia, 1970.

COLLEGE GPA AS A PREDICTOR OF TEACHER COMPETENCY: A NEW LOOK AT AN OLD QUESTION

TERRY L. JAMES
Westmar College
Le Mars, Iowa

WAYNE DUMAS
University of Missouri-Columbia

ABSTRACT

This study tests two hypotheses: (*a*) that the statistical relationship between academic success and success in student teaching is largely accounted for by those people in the lower grade point average categories; and (*b*) that four teacher competency ratings will be related to college grade point average in a pattern similar to that of global ratings of effectiveness. The study, which involved a sample of 170 University of Missouri secondary student teachers, resulted in the acceptance of both hypotheses with qualifications.

RECENT TRENDS IN TEACHER education have presented institutions with a perplexing paradox. The companion problems of teacher surplus and high enrollments in teacher education have brought about widespread institutional efforts toward more selective admission of primarily low-risk candidates to teacher education programs. Based upon five decades of teacher effectiveness studies (2-5) which lend modest support for college grade point average (GPA) as a predictor of teaching success, that criterion seems to be the one most frequently employed in selective admissions today (1).

Simultaneously, recent trends toward competency-based teacher education have increased pressures to focus attention on prospective teachers' mastery of specific completion competencies rather than upon the traditional, generalized global ratings of effectiveness used in previous studies. Proponents of competency-based criteria for teacher success often lead the chorus of criticism of the GPA basis for admissions, charging that if competency criteria were employed, probably no relationship would be found between college GPAs and success as a teacher. The present study tests the validity of this assumption.

A second assumption tested by this study is that the significant, though not high, positive correlations found in earlier studies between college GPA and measures of teaching effectiveness comprise a linear or uniform relationship. Many earlier studies operated statistically upon this assumption, concluding, in effect, that if knowledge of college GPA tells you a little bit about a C student's potential for success, it would tell you as much about a B student's potential and an A student's potential. However, it seemed probable to the authors that the relationship is far from uniform, and that while GPA might tell one a great deal about prospects for success by a 2.00 student, it might tell virtually nothing about the prosepcts for a student with a 2.80 average.

This study, then, was designed to test the following two hypotheses:

1. The statistical relationship between academic success, as measured by the cumulative college GPA, and success in student teaching, as measured by the six teacher effectiveness ratings used in this study, is accounted for by those people in the lower GPA categories. Figure 1 presents the graphic representation of this hypothesis.
2. Four competency ratings employed in this study will be related to college GPA in a pattern similar to that of the two global ratings of effectiveness. All patterns, in effect, will reflect Figure 1.

Procedure

Instrumentation and Data Collection

Two types of student teacher effectiveness ratings were used: global ratings and competency ratings in four categories.

Global Ratings: Two global ratings were used. These were the cooperating teacher's *Recommended Grade in*

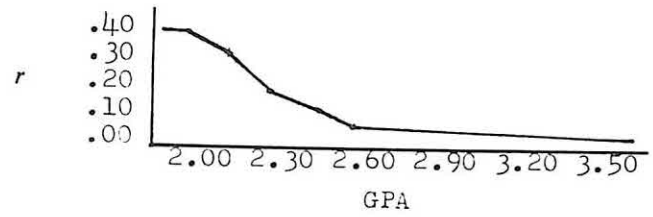


Figure 1.–Hypothesized relationship between academic success and student teaching effectiveness with a stepwise elimination of lower GPA categories from a hypothetical sample

Student Teaching (RGST) and a *Personal Impact Rating (PIR).*[1] The *PIR* utilized two hypothetical situations designed to commit the cooperating teacher in a very personal way. The first situation was concerned with the type of recommendation the cooperating teacher would give if the student teacher were being considered for a job in his/her school. Choices ranged from "cannot recommend" to "outstanding prospect, hire if available." The second situation was concerned with how strongly the cooperating teacher would recommend the enrollment of his/her child in a class to be taught by the former student teacher. The cooperating teacher could choose between the former student teacher and another beginning teacher about whom nothing was known, the rationale being that the beginning teacher about whom nothing is known would constitute the central or neutral position between positive and negative poles. Choices ranged from "definitely enroll with the former student teacher" to "definitely enroll with the other beginning teacher." The scores from the two situations were combined to produce the *PIR* rating for each student teacher.

The *PIR* test-retest reliability coefficient was .92, the highest reliability obtained for the criterion instruments used.

Teacher Competency Ratings: Four specific teacher roles or competency categories were identified, each considering the teacher as: (CR1) a stimulator; (CR2) a presenter; (CR3) an organizer; (CR4) a synthesizer. Each competency area was divided into two specific behavior components. A five-point continuum was used with descriptors at three points. Scores on the two subparts were combined for the rating on the particular role or competency.

The test-retest reliability coefficients ranged from .71 to .91.

Cumulative College Grade Point Averages: The cumulative college GPA was taken with the completion of the fifth semester or a minimum of 66 semester hours. Official student transcripts were used.

Subjects

Participants in this study were 170 secondary student teachers who had completed all of their work through the University of Missouri-Columbia. Student teaching was started and completed during either the fall semester 1972, or the first block, winter semester 1973.

The three subject areas of special education, vocational home economics, and vocational agriculture were excluded.

Analysis of Data

The 170 *S*s were categorized into seventeen different levels as determined by their cumulative GPA at the completion of semester five. GPA at the University of Missouri-Columbia is computed on a four-point system: C=2.00, B=3.00, A=4.00. These seventeen categorical levels were:

L1=all students
L2=2.00 and above
L3=2.10 and above
L4=2.20 and above
L5=2.30 and above
L6=2.40 and above
L7=2.50 and above
L8=2.60 and above
L9=2.70 and above
L10=2.80 and above
L11=2.90 and above
L12=3.00 and above
L13=3.10 and above
L14=3.20 and above
L15=3.30 and above
L16=3.40 and above
L17=3.50 and above.

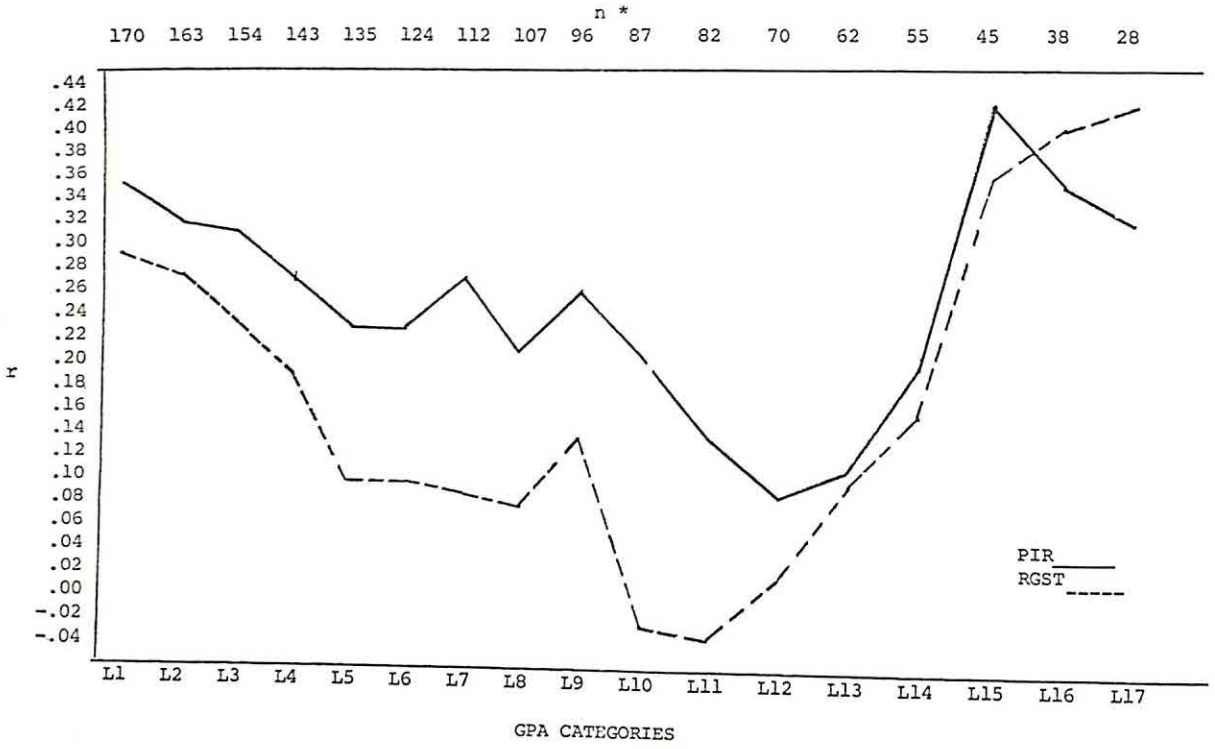
A Pearson *r* was computed between each criterion variable and each GPA category. This was accomplished for each variable through the use of a stepwise elimination procedure progressing from L1 to L17.

Results

As was anticipated from previous studies (2-4), correlation coefficients significant at the .001 level were found for the full sample between the cumulative GPA and both of the global ratings of student teacher effectiveness. Using the *Recommended Grade in Student Teaching (RGST)* as a criterion, the resulting coefficient was .29. Using the *Personal Impact Rating (PIR)*, the result was .36. The stepwise elimination procedure previously described was then employed, with the results graphically portrayed in Figure 2.

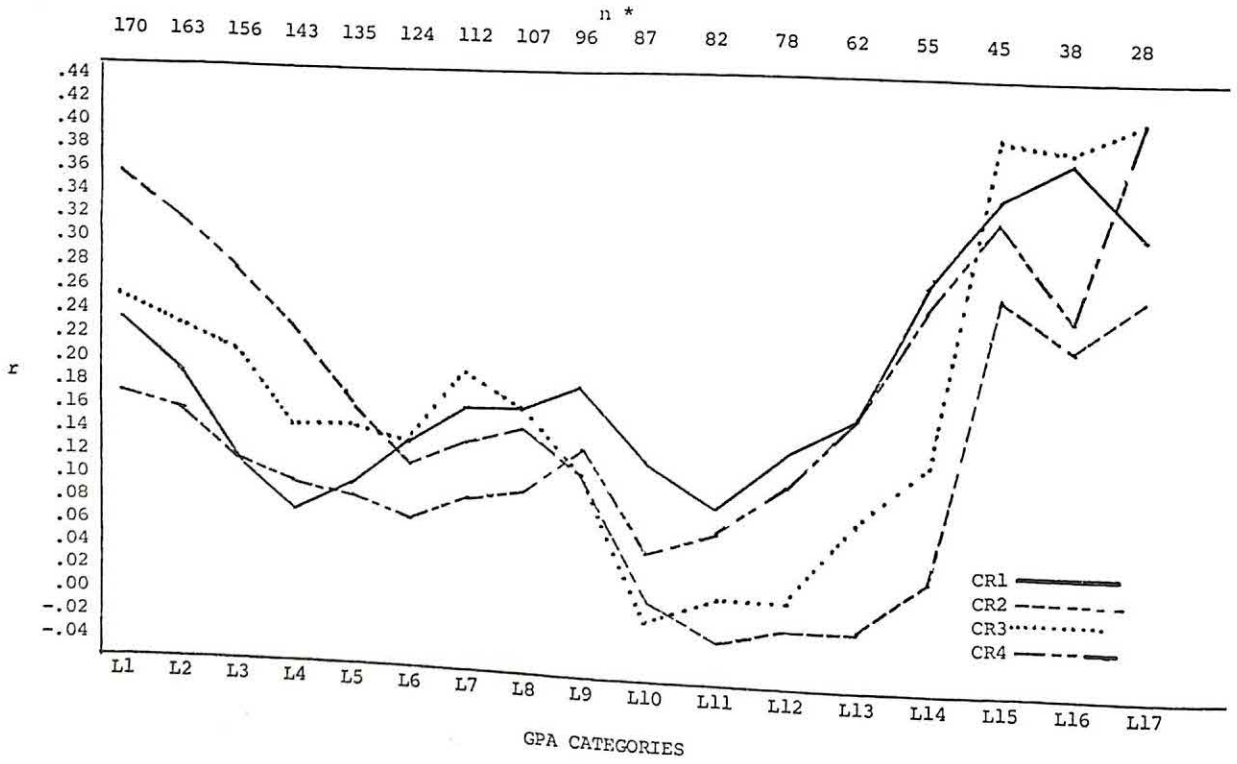
As student teachers with lower GPAs were systematically eliminated from the sample, the relationship between GPA and each of the two global criterion variables of student teacher effectiveness was rapidly diminished to statistical insignificance. Using the recommended grade criterion, the GPA became a statistically insignificant predictor of success (.05 level) when all students with GPAs below 2.30 were eliminated from the sample. Using the *PIR* as a criterion, GPA continued to serve as a statistically significant predictor of success at the .05 level until all students with GPAs below 2.90 had been eliminated from the sample.

A similar pattern emerges when the elimination procedure is employed with the four partial measures



* Subsample size after elimination of each GPA category

Figure 2.–Relationship between two global measures of student teaching effectiveness and cumulative GPA at semester five with a stepwise elimination of lower GPA categories from the sample



* Subsample size after elimination of each GPA category

Figure 3.–Relationship between four competency measures of student teaching effectiveness and cumulative GPA at semester five with a stepwise elimination of lower GPA categories from the sample

of effectiveness focusing upon specific competencies or roles of teachers (CR 1-4). Figure 3 demonstrates that all four measures of competency were initially correlated with GPA at the .05 level of significance, with correlations ranging from .36 (teacher as presenter) to .17 (teacher as synthesizer). GPA becomes a statistically insignificant predictor of all four criteria if all students with GPAs below 2.40 are excluded from the sample.

Based upon these findings, the authors are inclined to accept the first hypothesis, as graphically presented in Figure 1. In effect, GPA does indeed give some prediction of success or failure in teaching, regardless of the type of criterion one uses for success. But beyond a certain minimal level of demonstrated academic competence (approximately a 2.50 grade point level at the University of Missouri), GPA tells little if anything about a student's potential as a teacher. Apparently, other variables become more important.

An interesting and entirely unexpected qualification must be noted to the above inference. In both Figures 2 and 3, it will be observed that as the middle GPA group is also eliminated from the sample, all lines turn sharply upward again and five of the six criterion variables once again are correlated with the college GPA variable at levels exceeding the requirements for .05 significance, even with the considerably reduced sample size. The sixth variable, CR2 (teacher as presenter), follows the same pattern as the others but falls short of significance at the .05 level. Clearly, with this sample, the students who were highly talented academically were markedly more successful as student teachers than the upper-middle group. Finally, visual comparison of Figures 2 and 3 indicates that the second hypothesis is also correct. The pattern of relationship with GPA is essentially the same for all six criteria employed.

Discussion

The results of this study indicate that the use of college GPA as a selective admission criterion for teacher education may be useful and appropriate if used judiciously in combination with other variables. There is a positive relationship between GPA and both global and competency ratings of student teachers, but much of this relationship, with both types of criteria, is explained by the poor showing of very low GPA students in student teaching. The point at which the GPA becomes virtually worthless in selective admission would probably vary with the grading practices at various institutions. At the University of Missouri a grade point requirement exceeding 2.40 or 2.50 would probably serve no useful purpose in improving product quality.

NOTE

1. The *Personal Impact Rating* instrument was constructed Terry L. James.

REFERENCES

1. *A Survey of Current Practices in Selection and Retention of Students in Teacher Education,* (Monograph), University of Kentucky College of Education, Lexington, Ken., February 1972.
2. Kriner, H.L., "A Five-Year Study of Teacher College Admissions," *Educational Administrator and Supervisor,* 23:192-199, March 1937.
3. Mathis, C.; and Young Horn Park, "Some Factors Related to Success in Student Teaching," *Journal of Educational Research,* 58:420-422, May/June 1965.
4. Mead, A.R.; and Holly, C.E., "Forecasting Success in Practice Teaching," *Journal of Educational Psychology,* 7:495-497, October 1916.
5. Ott, V.K., "A Study of Some Techniques Used for Predicting the Success of Teachers," *Journal of Teacher Education,* 15: 67-71, 1964.

THE DELICATE ART OF TEACHER EVALUATION

WAYNE JONES
Stevens Point Area Schools
Stevens Point, Wisconsin

PAUL A. SOMMERS
Marshfield Clinic and
Medical Foundation
Marshfield, Wisconsin

ABSTRACT

In this study an attempt was made to provide an overview of applicable evaluation procedures which can be utilized by educational decision-makers responsible for the arduous task of teacher accountability. Specifically described are three different structures (models) of evaluation: systems; benefit-cost analysis; and experimental design. A discussion follows describing the types of measures to be applied and includes a review of the necessary teacher-student relationship under evaluation conditions.

HISTORICALLY, EDUCATIONAL ADMINISTRATORS have blamed inadequate resources, bureaucratic rule, political constraints, and a plethora of uncontrollable variables for their agencies' failure to attain educational goals. The returns for increasing investments in education have apparently been below the levels the consumer (the tax-paying public) has been led to expect. As a result, educational processes and outcomes are under critical investigation by the consumer, and the once authoritarian and learned work of the professional educator is regarded with unusual suspicion. In an effort to defend the utilization of dwindling resources, administrators have been searching for ways and means which would offer the necessary structure to permit systematic analysis and, it is hoped, solutions to their problems.

The problems which educational administrators continued to perceive as being most critical to the attainment of educational goals are fiscal in nature, i.e., additional personnel, experimental programs, federal projects, etc. Subsequently, educators have attempted to integrate planning-programming-budgeting systems into their educational systems for purposes of accurately dealing with such fiscal concerns. There have been many attempts to interpret the educational financial structure. Some approaches have been based on apparently "good" educational philosophies, while others are attempts to take concepts from successful business models. From the industrial world Leon Lessinger, past director of the Department of Health, Education and Welfare, introduced the concept of "educational program audit" based on the role of the certified public accountant; the acronym used in education is EPA (Educational Program Auditor).

In education an administrator's plight is inherently dependent upon the ability to motivate, guide, and evaluate instructional staff and their operational processes. Teacher evaluation has long been a problem confronting administrators, subsequently impeding any form of accurate accountability in the public education domain. To complicate matters, the inferred relationship concerning appropriate and efficient teaching practices must deal with a very important variable, the student. And when one wishes to investigate the relationship between teacher and student, a multidimensional task is inevitable. Concomitantly, one must take into account such variables as school-related student attributes, non-school student-related attributes, program and service variables, student performance variables, post-school adjustment variables, and many, many more.

Given this set of critical and complex circumstances, the administrator is required to implement a sound, scientifically based instructional evaluation system, complemented by an equally appropriate data collecting and monitoring system. In a majority of cases, expert outside assistance is necessary for at least evaluation design purposes. However, due to lack of finances, the responsibility of designing such an instructional evaluation system is often routinely delegated to program administrators.

In many cases these administrators begin to feel literally trapped. The "trapped administrators," according to Campbell (5), are those who "have so committed themselves to the efficacy of the reform that they cannot affort honest evaluation." A contrast is made when those administrators who "initially justified the need for reform on the basis of importance of the problem, not the certainty of their answer, are committed to going on to other potential solutions if the one first tried fails" (5). ("Reform" here is used by Campbell to indicate a commitment to a new program.) Because of these facts the educational administrator creates a system that forces a high probability of making a variety of instructionally inappropriate program

choices. Examples of these poor decisions are found in curriculum innovations, teaching methods, staffing models, and pupil management approaches.

In an attempt to develop a systematic procedure for dealing with the task of making appropriate and efficient decisions, the following presentation will suggest certain means of carrying out teacher evaluation and effectively utilizing the data produced.

Methods employed in the development of an instructional evaluation system (evaluation of teachers) can be generalized to any personnel evaluation problem. Teachers were chosen as subjects of this paper on evaluation because: (*a*) their actions are often considered more visable than the administrator and, therefore, subject more often to criticism; (*b*) many administrative policies are initiated through teachers; and (*c*) traditionally administrators are responsible for the process of teacher evaluation. Few texts in educational supervision and administration provide extensive discussion of teacher evaluation methods; however, the importance of teacher evaluation is invariably mentioned. Lindley (18) states that "evaluation of teaching is not only desirable but quite necessary and even inevitable." Others tend to support this statement.

Teacher Evaluation

A good deal of varied opinion exists regarding the form teacher evaluation should take. Gage (14) recommends that the complex behavior of teaching be broken down to measure "micro-effectiveness." He says:

> Rather than seek criteria for overall effectiveness of teachers in the many, varied facets of their roles, we may have better success with criteria of effectiveness in small, specifically defined aspects of the role; if such laws could be developed, they might eventually be combined . . . to account for the actual behavior and effectiveness of teachers with pupils under genuine classroom conditions.

Saadeh (22) states that:

> Viewing the whole phenomenon of teaching effectiveness in terms of its parts seems to ignore the necessity of treating the teacher–learning act as a totality.

Evaluation Designs

An attempt will be made to discuss the subject of teacher evaluation from the perspective that the only purpose for evaluation is to supply information upon which more appropriate and efficient instructional-based decisions can be based. Data from evaluative processes may be applied to decisions in two ways. The first is to indicate that a decision should be made. The second is to support decisions already made (23). Both may be found in evaluation designs that are either *post-hoc* or *prearranged*. Figure 1 compares the two designs in terms of the sequences of program events. In the post-hoc example, data are collected, evaluation designed, data analyzed, and decisions made. In the prearranged situation, evaluation is designed, program initiated, data collected and analyzed, and the decision is made.

In the post-hoc situation, data are gathered with unavoidable bias. Unless the data collection process takes place under a prearranged design, "the study must be classified as a false experiment" (4). Examples of post-hoc designs can be found in the Westinghouse, Ohio (26) study of Headstart and the Dentler (10) review of the More Effective Schools program in New York.

Contrasting the post-hoc approach, prearranged designs of data collection require that the educational administrator has anticipated a decision point or situation and has designed the evaluation procedure to facilitate an enlightened decision based on the appropriate data.

The forces or motives dictating the types of decisions determine, to a large extent, the form the evaluation is to take. The resultant data then give evidence of the action the administrators should take and enable the administrator to facilitate this action after the decision.

In order to provide educational decision-makers with a basic understanding of appropriate evaluation design considerations, an illustration and discussion of various structures (forms), with accompanying measurement techniques, in which teacher evaluation can take place will follow. Evaluation structures to be discussed include a system's approach to evaluation, a benefit-cost analysis approach, and an experimental design approach. Although these three approaches far from represent the totality of procedures available within the evaluation domain, they do appropriately serve as viable alternatives within the arena of teacher evaluation.

Structures of Evaluation

The public school administrator has traditionally been responsible for processes of planning, organizing, allocating resources, staffing, coordinating, controlling, and evaluating. Too often administrative structures evolve into mechanisms to facilitate the functioning of these processes individually. A balanced budget or complete staff, however, does not necessarily imply that educational goals are being met. Information derived from the analysis of individual processes cannot be used to infer institutional approximation of educational goals. In order

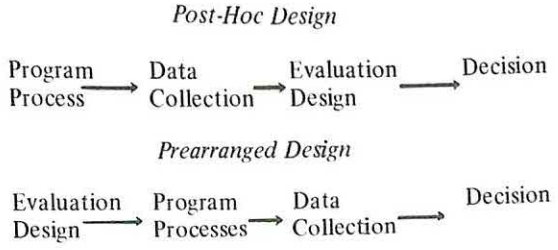


Figure 1.–Comparison of post-hoc and prearranged designs

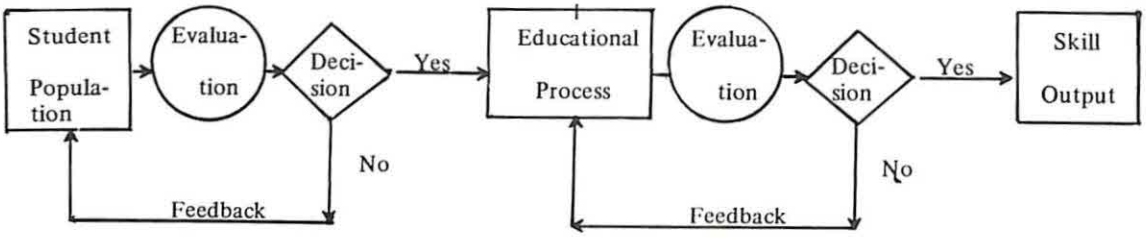


Figure 2.—A model educational system

to determine goal approximation with some reliability, the administrative processes must be superceded by a system (17). It is also true that goal achievement or teaching processes of teachers are best analyzed within the context of the total educational system rather than the individual teaching situation.

Systems

The evaluation of a teacher and the interpretation of information resulting from such an evaluation require a great amount of time and effort on the part of those involved. Usually the school administrator goes through a process of determining the outcome of an evaluation by his perception of the data before and after collection. The evaluation process is, no matter how sophisticated, limited in usefulness when not conducted and described as an integral part of a system. Without the structure of a system, the administrator is not entirely able to visualize the impact of the information derived from the teacher evaluation process. As a result, the administrator may react inappropriately toward a teacher's behavior or underestimate the value of teacher evaluation.

Figure 2 illustrates a system in terms of levels, flows, and decision points (13). The system of Figure 2 represents a student population that has been evaluated as having minimal skills for entry into a specific educational program such as high school physics. The evaluation processes serve the decision points as well as provide information to the administrator on the effectiveness of the system. To assess the teachers' effectiveness in this system, the evaluator must consider the nature of the system itself, which basically includes the following variables: student entry skills; variance of skills among students; total number of students; skills of the teachers; and the school's financial investment.

Benefit-Cost Analysis

A number of the major decisions made by administrators involve the allocation of funds. And "the relationships between application of resources to a particular program and attainment of objectives can be determined by benefit-cost analysis" (9). Benefit-cost analysis has been defined as the ratio of the present value of future benefits to the present value of future costs. In a decision-making situation, one must take into consideration the fact that present resources are always more valuable than future resources. The school administrator evaluating teacher performance does so usually with present costs paramount in mind.

Cost and benefit of vocational programs may be determined through wages earned and taxes paid by past students. Cost and benefit analysis in an elementary school might be determined on the basis of program comparisons and, therefore, relate teacher evaluation to program cost. Such a relationship permits a more valid evaluation of teachers by weighting programs according to resources. A program receiving a great deal of support should show more benefit than one with a smaller allocation of resources—all other variables being equal. A teacher appearing to have greater results with a small resource allocation will usually receive a higher commendation than a teacher receiving the same results with a larger allocation of resources.

By evaluating a teacher on the basis of benefit-cost analysis, one may determine that the achievement (or lack of achievement) of objectives may be costing more through one program than another or through one teacher than another. However, *cost* does not necessarily represent material considerations but may be studied in terms of time required to complete an objective (time of educators or students) or the total number of individuals involved (number of educators or students). Consequently, the benefit-cost analysis approach requires that a quantity (variable) be designated and the appropriateness and efficiency of the use of the quantity be analyzed and related to the program (objectives). The effects of the variable on the program objectives are determined through the most appropriate evaluation or experimental designs.

Experimental Design

In order to provide structure to the procedure of teacher evaluation, experimental design considerations are essential. The designs are presented in order of scientific experimental strengths—weakest to strongest. The first four are excerpted from *Experimental and Quasi Experimental Designs for Research* by Campbell and Stanley (6).

1. *Interrupted Time-Series Design*—In this setting the comparison base is the record of previous years.

The usual mode of application is a casual version of a quasi-experimental design, the one-group pre-test/post-test design. This is used where no control group is possible.

The interrupted time-series design utilizes previous years or months of performance of a teacher or group of teachers by plotting identified performance against the selected criterion. This design ignores all the variables and their variance over the time in question. It does, however, generate questions and therefore other studies and theories.

* * * * *

2. *Control Series Design*—The most common of such designs is the non-equivalent control group pre-test/post-test design, in which for each of two representative groups, pre-test and post-test measures are taken and one receives the treatment.

The purpose for this design is to analyze the variance in group performance. When applied to teacher evaluations, this design may be utilized for students or groups of teachers who are selected in a representative manner.

* * * * *

3. *Regression Discontinuity Design*—If randomization is not politically feasible or morally justifiable in a given setting, there is a powerful quasi-experimental design available that allows the scarce good to be given to the most needy or the most deserving. All it requires is strict and orderly attention to the priority dimension.

In a situation where the introduction of a program or teaching method is being piloted on a group of children or teachers, initiation may be by those who fall above or below some decision point on an achievement measure, rating, or performance continuum. For example, if an administrator wishes to pilot the use of an innovative teaching technique under optimal conditions, he may rate teachers and/or students according to some criterion and choose those who fall above a criterion "cut-off" point.

After an appropriate predetermined period of time, the criterion is again applied and the administrator takes note of the difference (if any) in rating at the decision point.

* * * * *

4. *Randomized Control Group Experiments*—Randomized experiments tend to be limited to the laboratory and/or agricultural experiment stations. But this certainly need not be so. The randomized population may be persons, families, precincts, or administrative units.

This experimental design is sometimes called the *true* experimental design. Although large groups are desirable, representative selection is possible within a single school, depending on the evaluation design and the level of confidence (significance) the administrator believes is necessary for a decision.

* * * * *

5. *Single Subject Design*—The use of replication with comparisions of an individual's behavior rate changes before and after experimental intervention. Because the experiment is not dependent on numbers, intervening variables may be controlled more accurately than in any of the other designs. The subject becomes his own control.

Single subject research designs are not based in statistical generality, and subsequently are very useful in individual teacher evaluation. As Sidman (24) says:

> Once the administrator has pointed out those features of teacher performance with which he is particularly concerned, . . . direct replication of the teaching activities may be accomplished either by performing the experiment again with new subjects or by making repeated observations on the same subjects under each of several evaluation conditions.

* * * * *

6. *Multivariate Analysis*—Given the multidimensional task of having to concomitantly take into account such variables as student attributes, non-school environmental variables, program and service variables, student performance variables, and post-school adjustment variables necessitates utilization of a multidimensional approach termed "multivariate analysi (28). Multiple linear regression analysis, a form of multivariate analysis, was selected by Sommers and Joiner (27) for purposes of conducting research when a variety of behaviorally oriented variables were being investigate in a study of the "disadvantaged." They state:

> It is assumed that performance or behavior is subject to the influence of more than one variable or condition at a time and that adequate explanations involve more than a single variable or condition. But, if several variables are proposed as being relevant to performance, it becomes necessary to measure both the influence of the variables on the behavior we are attempting to explain and their influence upon each other.

Multivariate analysis allows the evaluator to reflect complexities in the evaluation paradigm. The power of prediction as an intellectual tool resides in the fact that it enables one to rigorously test the adequacy of various theoretical evaluation models that might be proposed.

The various designs were described in order to suggest that techniques of teacher evaluation should have as their basis an acceptable and appropriate evaluation design with pre-set standards rather than some after-the-fact rationale. In each of the designs, the variables must be defined in a quantifiable format and specifically relatable for purposes of administrative decision-making.

Evaluation Measures

Cognition, Affect, and Performance

Almost all measures applied to the domain of teacher evaluation attempt to describe content areas of cognition, affect, and/or performance. In the cognitive realm, mental ability, knowledge of subject matter, educational level, and verbal ability are common areas of measurement. Although some investigators (16) have shown that the intelligence of teachers is highly correlated with student achievement, others (15) have demonstrated that little is known concerning the relationship between cognitive abilities of teachers and variables commonly associated with student achievement.

Similar complexities also exist in the relationship of student accomplishment and the teacher's command of subject matter. Generally, knowledge of subject matter is considered essential to teaching; however, when one considers the goals usually set for the early years of elementary education (reading, spelling, arithmetic, etc.), knowledge of subject matter becomes less important than such categories of skills as teaching methods, curriculum design, behavior management, etc. It has been documented that sufficient ability exists to equip a teacher with information related to subject matter (history, chemistry, mathematics, etc.). Consequently, educators should turn their efforts to the more difficult task of training the teacher to teach. Verbal ability of a teacher, usually demonstrated on a performance level, has been shown to have high correlations with pupil achievement (8). In contrast, a teacher's educational training is often considered as an inaccurate indicator of teacher effectiveness (11).

Affect (i.e., attitudes, interest, sense of humor, etc.) is usually defined by the evaluator and/or instrument that measures it (7). Difficulty exists in obtaining valid measures of affect, and teacher evaluation based on affective measures may result in information in conflict with data derived from measures having their bases in more concrete and less subjective areas.

Complexity also exists in relation to the validity of performance-based measures. Barro (2) says that "no program of performance measure alone, no matter how comprehensive or sophisticated, is sufficient to establish accountability." One of the major problems in the use of performance measures is the divergent definition of performance. The Skinnerian (25) definition, involving counting and recording only directly observable behavior, imposes rigid constraints on the evaluation procedure in performance areas. However, such constraint provides extremely reliable and valid data on which to base decisions.

The four most common measures used in the evaluation of teachers are ratings, achievement scores, categorization, and event counting. Each of these may be applied by an observer, student, and/or teacher. Table 1 illustrates the various measures and their applicability to each of these three groups of "users." The table shows that the use of observers occurs more often than the use of students; ratings are shown to be used more often than the other scales.

Table 1.–Teacher Evaluation Measures and Users

Users	Ratings	Achievement	Categories	Event Counting
Observer	1	1	1	1
Student	1	0	0	0
Teacher (self)	1	0	0	1

1–indicates use of a scale
0–indicates little or no use of a scale

Rating Scales

Ratings and rating scales are the most commonly used measures. Ralunowitz and Travers (20) suggest a good rating scale should: (*a*) define with precision several points on each scale; (*b*) restrict each scale to well-defined and observable behavior; (*c*) vary the end of the scale (where several are used) which represents "good"; and (*d*) avoid the use of words such as "average." The two most useful rating scales are *interval* and *ratio*. Almost all the usual statistical measures are applicable to the interval scale unless knowledge of a "true" zero is required (30). The ratio scale requires the existance of true zero. All the statistical measures applicable to the interval scale apply to the ratio scale, as well as geometric mean, coefficient of variation, and logarithmic scales. The behavior of a teacher may be rated on a direct magnitude estimation scale in order to develop a data basis for statistical manipulations.

Achievement Measures and Categorizations

Achievement measures are indirect ways of evaluating teacher effectiveness (2) and attract sufficient criticism from inferences often drawn from their application. While ratings and categories are criticized for problems arising from inferring causality from correlations, achievement measures tend to suffer from inappropriate inferences drawn from normalized sampling distributions.

Categorizations, as in the "Variant-Flanders Interaction Analysis" (6:21), consist of attempts to describe and categorize teacher behavior. Observations are made during "encounters" or probes set at specific time intervals. The trained observer, which may be an administrator or teacher viewing a video-tape, checks the various categories of behavior occurring in a specific encounter. The categories are often arranged in groups, such as student-centered, subject-centered, student behavior, and teacher behavior. The total number of times a behavior is observed indicates the percentage of time spend in exhibiting the various behaviors (as in Openshaw (19), 80% teacher verbal behavior to 20% student verbal behavior).

Sorenson and Gross (29) used a categorical process which proceeded from the assumption that

> a teacher may be said to be "good" only when he satisfies someone's expectations, that people differ in what they expect from teachers, and that a scheme for evaluating teachers and for predicting their effectiveness must take into account categories related to instructional objectives, methods of instruction and teacher relationships with pupils.

Rosenshine (21) divided instruments used for teacher evaluation into *category systems* and *rating systems*, the difference between the two being the amount of *inference* required of the observer. Inference here refers to the process intervening between the objective data seen or heard and the coding of those data on an observational instrument. Categorization requires the least inference, and the encounters the teacher has are counted (e.g., teacher asks evaluative question) in a manner similar to event counting. Rosenshine's ratings require the formation of inferences and are used to rate such qualities as enthusiasm and vigor.

Event Counting

Event counting is similar to categorization; however, it generally is used when a goal has been defined in terms of a behavioral objective. This form of evaluation is an integral part of meeting program objectives. Assuming that the rates or frequencies resulting from the application of behavior modification techniques are interpretable as measures of performance, and assuming that one believes it is possible and necessary to arrange contingencies and objectives for a teacher, then counting can be considered a teacher evaluation measure. One of the most appealing characteristics of event counting is that by setting specific behavioral objectives for the teacher, one eliminates the need for an inference to be drawn from the data collected.

Applying the Measures

Users

The use of a trained observer is a desirable method of applying the four measures. A trained observer provides reliability and objectivity for the measure being utilized. In some cases, the extensive experience required for reliable application of an instrument (e.g., Variant-Flanders) prohibits extensive use of the measure. An observer may be used occasionally to establish "inter-rater reliability" in ratings, categorization, and event counting.

Teacher (self) application of any of measures for teacher evaluation has the problem of relative objectivity. Generally, when the teacher rates himself/herself or uses a matching instrument, the resulting scores lean in the direction of the teacher's self concept and are relatively high when compared with the scores of an observer. The teacher also has some problems with reliability in event counting if contigencies are not established in a pre-set format prior to the counting process.

The student's role in evaluation is often evidenced through ratings and rating scales. The Educational Feedback Center (EFC) at Western Michigan University is a system based on student ratings. A profile is produced which represents the average student's reactions to questions believed to be related to teacher effectiveness. This process, as well as most processes involving student ratings, has not been determined applicable during the early elementary school years (EFC usually pertains to children in grades 7-12).

An example of student perceptions of teachers using the "Teacher Image Questionnaire" from Western Michigan was compiled by William Coats (7), who did a factor analysis of 42,810 student responses in which

> a single factor, labeled teacher "charisma," was found to account for 61.5% of the variance in test items. Five other factors accounted for the balance. It was concluded that teacher charisma is probably a factor of teacher effectiveness, but that student ratings would best be used as only one part of a total evaluation . . .

Criterion Behavior

When any evaluation is initiated it is necessary to define "criterion behavior" or the reason for the need to evaluate. By utilizing the information found in Table 1, each user must decide on the criterion behavior or expected result to be evaluated using the various measures. In every case the measure is applied to some observable behavior. The inferences made from the observed behavior may lie on a continuum between valid and invalid. The most valid measure is event counting when the criterion is simply a predetermined rate of the observed behavior. Reduced validity develops as the observed behavior is separated from criterion behavior by inferences and/or theories.

Earlier mention was made that the functional value of evaluation can be evidenced by the number of appropriate decisions. The type of decision required will determine the purpose and form of the evaluation. Relating this to the previous discussion of criterion behavior, the evaluation process and the quality of the decision is dependent on the appropriateness and the representativeness of the criterion chosen (20).

Bolvin's (3) investigation of teacher performance in the area of prescription writing for various students illustrates how various criteria were related to prescription writing as a measure of effectiveness. Of interest in the Bolvin study was the rationale for choosing to evaluate the "prescription" as a reflection of teacher effectiveness. Prescriptions were chosen as one aspect of teacher activity that leaves a record. The evaluation was based on the teachers' criterion for writing a specific prescription and the perceived constraints (i.e., time, variety of materials, etc.), a key point being the critical need to identify and monitor the tangible evidence of teacher performance.

Summary

Teacher evaluation can be thought of as either an *informal* or *formal* process. The reason necessitating teacher evaluation will primarily determine the form it is to take and will guide the administrator in the selection of instrumentation and associated implications. Typically, teacher evaluation is informal. Simple rating instruments or achievement measures are utilized by an administrator or supervisor. Decisions made as a result of such informal evaluation seem to fall into two general categories. First, during teacher probation or early employment periods, results of evaluation reflect on decisions to recommend continued employment. The probability that anything but general incompetency or problems of affect would result in dismissal of a tenured teacher is quite low. When such is the case, teacher evaluation methods would most likely be used to "justify a decision" already arrived at through other processes.

A second type of decision typically resulting from the informal evaluation process is on a personal level for the teacher being evaluated. The results of such an evaluation may assist the teacher in deciding on changes in teaching methods and approaches.

The term *formal* is meant to imply that the evaluation process is set within a defined format. The discussion presented in this paper, with the teacher evaluation process in a structure, has outlined the reasons why an administrator should consider the formal approach. However, the informal encounter with a teacher, for evaluation purposes, will probably continue to occur more often than the formal evaluation. The reinforcing elements necessary for developing an appropriate teacher evaluation procedure and projecting data for such decisions is all but missing in the day-to-day world of most administrators. But, due to the impact of legislative decisions eminating from a national level and directly affecting every corner of public education, there is a critical need for educational administrators and teachers to self-impose a refined system of accountability before it is done for them by unaffective, non-educational governmental audit and accounting agencies.

REFERENCES

1. Andrew, G.; and Moir, R., *Information-Decision System in Education,* Peacock, Itasca, Ill., 1970
2. Barro, S., "An Approach to Developing Accountability Measures for Public Schools," *Phi Delta Kappan,* 52:231-235, December 1970.
3. Bolvin, J.O., "Evaluating Teacher Functions," paper presented at the Meeting of the American Educational Research Association, University of Pittsburgh Learning Research and Development Center, 1967.
4. Campbell, D.T., "Factors Relevant to the Validity of Experiments in Social Settings," *Psychological Bulletin,* 54:297-312, 1957.
5. Campbell, D.T., "From Description to Experimentation: Interpreting Trends–Quasi Experiments," in H. Chester (ed.), *Problems in Measuring Change,* University of Wisconsin Press, Madison, Wisc., 1963, pp. 212-243.
6. Campbell, D.T.; and Stanley, J.C., "Experimental and Quasi Experimental Designs for Research," Rand McNally, *Chicago,* 1963.
7. Coats, W.D., "Student Perceptions of Teachers: A Factor Analytic Study," paper presented at the American Educational Research Association Convention, Minneapolis, Minn., American Educational Research Association, Washington, D.C., 1970.
8. Coleman, J. *et al,* "Equality in Educational Opportunity," U.S. Office of Education - HEW, Government Printing Office, Washington, D.C., 1966.
9. Davie, B.F., "Using Benefit - Cost Analysis in Planning and Evaluating Vocational Education, ERIC - ED - 016-077, 1965.
10. Dentler, R., "Urban Eyewash: A Review of Title I Year II," *Urban Review,* September 1969.
11. DeVane, L.M., "The Qualities and Qualifications of Excellent High School Teachers," unpublished doctoral dissertation, Florida State University, 1961, p. 60.
12. Elsbree, W.S., *et al., Elementary School Administration and Supervision,* American Book Co., New York, 1967, p. 166.
13. Forrester, J., *Principles of Systems,* Wright, Allen, Cambridge, Mass., 1968.
14. Gage, N., "An Analytical Approach to Research on Instructional Methods," *Journal of Experimental Education,* 37:119-125, 1963.
15. Getzels, J.W.; and Jackson, P.W.; "The Teacher's Personality and Characteristics, in N.L. Gage (ed.), " *Handbook of Research on Teaching,* Rand McNally, Chicago, 1963, p. 571.
16. Jones, M., "Analysis of Certain Aspects of Teaching Ability," *Journal of Experimental Education,* 35:103-108, 1956.
17. Knezevich, S., "Administrative Technology and the School Executive," American Association of School Administrations, 1969.
18. Lindley, J.S., "The Cooperative Research Program Contribution and Next Steps," *Phi Delta Kappan,* 23:231-236, March 1962.
19. Openshaw, M.K. *et al.,* "The Development of a Taxonomy for the Classification of Teacher Classroom Behavior," Ohio State University Research Foundation, Columbus, Ohio, ERIC-ED-010-167, 1967.
20. Ralunowitz, W.; and Travers, R.M., "Problems of Defining and Assessing Teacher Effectiveness," *Educational Theory,* 3:212, July 1953.
21. Roseshine, B., "Evaluation of Classroom Instruction," *Review of Educational Research,* Vol. 40, No. 2, 1970.
22. Saadeh, I.Q., "Teacher Effectiveness of Classroom Efficiency: A New Direction in the Evaluation of Teaching," *Journal of Teacher Education,* 21:73-91, 1970.
23. St. John, S., "Segregation and Minority Group Performance," *Review of Educational Research,* 40:111-129, 1969.
24. Sidman, M., *Tactics of Scientific Research,* Basic Books, New York, 1960.
25. Skinner, B.F., *The Behavior of Organisms: An Experimental Analysis,* Appleton Century Crofts, New York, 1938.
26. Smith, M.; and Bissell, J., "Report Analysis: The Impact of Headstart," *Harvard Educational Review,* Vol. 40, No. 1, 1970.
27. Sommers, P.A.; and Joiner, L.M., "Kinesio-Perpetual Abilities of Predictiors of Race: A Study of the Disadvantaged," *Negro Educational Review,* October 1970.
28. Sommers, P.A., "An Inferential Evaluation Model," *Journal of Educational Technology,* May 1973.
29. Sorenson, G.; and Gross, C.F., "Teacher Appraisal–A Matching Process," ERIC - ED - 016-299, 1967.
30. Stevens, S.S., *Handbook of Experimental Psychology,* Wiley, New York, 1951.
31. Weber, M., *The Theory of Social and Economic Organizations* (trans. by A.M. Henderson and T. Parsons), Free Press, New York, 1947.

LIMITATIONS OF ANALYSIS OF COVARIANCE ON INTACT GROUP QUASI-EXPERIMENTAL DESIGNS

PAUL A. GAMES
The Pennsylvania State University

ABSTRACT

Multiple regression models are used to demonstrate that every organismic variable is to some extent a proxy for every pertinent missing organismic variable. For analysis of covariance, clear assessment of treatment effects is possible only when the treatment vector(s) is(are) kept orthogonal to *all* organismic variables by random assignment of subjects. The "adjusted treatment effects" of covariance analysis on quasi-experimental designs include effects resulting from differences in the adjusted means of the treatment groups on pertinent organismic variables—*both* those used as covariates and others that are missing from the analysis. Only if the adjusted treatment means do not differ in *any* of the organismic variables that are pertinent for predicting the criterion would the assessment of treatment effects be proper.

COVARIANCE AS A TECHNIQUE that may be used to correct for confounding of organismic variables when subjects have not been randomly assigned to treatments was presented by McNemar (10:413-414) and Ferguson (5:326). Organismic variables are variables that may be obtained by measurement of subjects, but that are not assignable to subjects. Mental age, sex, ability measures, personality measures, past education, etc., are among the hundreds of interrelated organismic variables. Organismic variables may be contrasted to manipulatable variables, or treatments, that may be assigned to any subject.

After the above texts were in press, Lord (8, 9) and Cronbach and Furby (1) questioned the use of covariance on groups that initially differ in one or more organismic variables. Evans and Anastasio (4) distinguished three logically different uses of covariance: *Use one*—where subjects have been randomly assigned to groups; *Use two*—where intact groups are assigned to treatments and covariance is used to "adjust" for differences between the group means on observed organismic variables; and *Use three*—where the differences in the covariate means are the result of different treatments (as when final trial learning measures are used as covariates on retention scores). Evans and Anastasio argue against the last application, but accept the second application if the covariate is measured before the treatments are administered and the intact groups are randomly assigned to treatments. The present article argues that *use two* is also unlikely to lead to interpretable results.

Multiple regression (MR) is a general data analysis technique that includes all of analysis of variance (ANOVA) and analysis of covariance (ANCOVA) as special cases (6). For simplicity, we may use vectors of deviation scores, $\underset{\sim}{x}_j = \underset{\sim}{X}_j - \underset{\sim}{\overline{X}}_j$ as predictors, and $y_i = \overline{Y} - \overline{Y}$ as the criterion with means computed over all subjects. Or equivalently, we take the means as zero, with no loss of generality.

The general model for three predictors is:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i$$

or, in vector terms,

$$\underset{\sim}{y} = \beta_1 \underset{\sim}{x}_1 + \beta_2 \underset{\sim}{x}_2 + \beta_3 \underset{\sim}{x}_3 + \underset{\sim}{e}$$

The general test for whether any given variable makes a nonchance contribution to predictive accuracy (given the other variables are used) is a test of H_o:

$$\beta_j = 0 \text{ by } F_o = \frac{R^2_{Y.1\ \ldots p} - R^2_{Y.1\ \ldots j-1, j+1, \ldots p}}{(1 - R^2_{Y.1\ \ldots j\ \ldots p}) / (N - p - 1)}$$

where p = the total number of predictors. Since SS_Y = SS (regression $1 \ldots p$) + SS_E and SS (regression $1 \ldots p$) = $R^2_{Y.1\ \ldots p} SS_Y$, the same statistic may be formulated as

$$F_o = \frac{\text{SS (add. reg. } j)}{SS_E / (N - p - 1)}$$

where SS (add. reg. j) = SS (reg. $1 \ldots p$) - SS (reg. $1 \ldots j-1$, $j+1 \ldots p$). Thus, in a three-variable problem, to test H_o:

$$\beta_2 = 0, F_o = \frac{\text{SS (add. reg. 2)}}{SS_E / (N - 3 - 1)}$$

is compared to $F(a, 1, N - p - 1)$ where SS (add. reg. 2) - SS (reg. 123) - SS (reg. 13). This test requires the usual assumptions of classical MR (7:95); however, $E(b_j) = \beta_j$ even when the normality and heterogeneity assumptions have been violated.

In MR it is important to distinguish between orthogonal and nonorthogonal vectors. Two deviation score vectors $\underset{\sim}{x}_1$ and $\underset{\sim}{x}_2$ are orthogonal when the covariance between them, δ_{12}, is zero. The condition of orthogonality creates great simplicity in MR. If we have a set of three mutually orthogonal predictors,

$$\rho^2_{y.123} = \rho^2_{Y1} + \rho^2_{Y2} + \rho^2_{Y3}$$

and we may determine the independent contribution of each predictor to $\rho^2_{y.123}$

When orthogonality is absent, no independent contribution can be identified with the individual predictors (2). For the nonorthogonal case, β_3 is the bivariate regression slope that exists between y and the residual vector of x_3 after x_1 and x_2 have been partialled out. That is, β_3 is the bivariate regression coefficient of a scatterdiagram, with y on the vertical axis and $\tilde{\underset{\sim}{x}}_{3.12}$ on the base axis; $\tilde{\underset{\sim}{x}}_3 = \underset{\sim}{x}_3 - \hat{\underset{\sim}{x}}_{3.12}$ where $\hat{\underset{\sim}{x}}_{3.12}$ is the predicted x_3 from the multiple regression of x_1 and x_2. Thus, β_3 is influenced not only by the yx_3 relationship, but also by both x_1 and x_2.

It is well known that for the nonorthogonal case, β_3 may be drastically changed if either x_1 or x_2 is dropped as a predictor. For orthogonal vectors, however,

$$\tilde{\underset{\sim}{x}}_{3.12} = \underset{\sim}{x}_3 = \tilde{\underset{\sim}{x}}_{3.1} = \tilde{\underset{\sim}{x}}_{3.2}$$

and the β_3 value remains the same whether $\underset{\sim}{x}_1$ and $\underset{\sim}{x}_2$ are included in the equation or not. The stability, simplicity, and clarity of the orthogonal case are greatly to be desired, but are rarely achieved with organismic variables as predictors. Organismic variables are "intrinsically nonorthogonal" in that any such variable has non-zero covariances with a large set of other organismic variables. The methodology of factor analysis has been created in an effort to make conceptual sense of matrices of such covariances.

In contrast, treatment main effects may be conceived of as "intrinsically orthogonal" to organismic variables in that any subject may be assigned to any given treatment level. With proper random assignment of subjects in experiments, there is no tendency for subjects with high x_1 values to be in A_1 and subjects with low x_1 values in A_2. For simplicity, consider two independent groups of n cases ($N = 2n$) so that the treatment may be represented by a single vector $\underset{\sim}{A}$, where each subject in the experimental group is assigned as +1 and each subject in the control group is assigned as -1. The $\underset{\sim}{A}$ vector has a mean of zero, and may be treated as a deviation score vector. With random assignment of subjects to groups, then $\delta_{Ax_j} = 0$ for any $\underset{\sim}{x}_j$, and $\underset{\sim}{A}$ is orthogonal to all possible organismic variables.

The ANOVA model for this simple experiment may be formulated as $\underset{\sim}{y} = \beta_a \underset{\sim}{A} + \underset{\sim}{e}$. Here $\beta_a = .5(\mu_{YE} - \mu_{YC})$ where E and C represent the experimental and control groups respectively. β_a directly reflects the amount of treatment effect. The test of H_o: $\beta_a = 0$ by $F = \text{SS (reg. } A)/ \text{MS}_E$ is an algebraic equivalent of the usual t-test of means of two independent groups.

To obtain an ANCOVA model for this situation, we merely add additional organismic variables as predictors. In *use one*, these vectors are orthogonal to the A vector, hence β_a is unaffected by the addition of the new variables, and is exactly the same as in the ANOVA analysis above. The model when two variables are used as covariates is:

$$\underset{\sim}{y} = \beta_1 \underset{\sim}{x}_1 + \beta_2 \underset{\sim}{x}_2 + \beta_a \underset{\sim}{A} + \underset{\sim}{e}.$$

In sample data, chance variations from orthogonality may occur, and corresponding minor variations in b_a may result; *however, the same parameter is being estimated.* The prime impact of the use of $\underset{\sim}{x}_1$ and $\underset{\sim}{x}_2$ as covariates is that they should reduce MS_E if they are indeed effective predictors of y. Now $\text{SS}_E = \text{SS}_Y - \text{SS (reg. } 12A)$ so that if SS (reg. $12A$) is much larger than SS (reg. A), a substantial increase in power will occur. The major role of ANCOVA is to increase power over that from ANOVA. The treatment effect, β_a, is not influenced by whether the organismic variables are used or not in *use* one.

In contrast to the above situation, consider what happens in *use two* when one intact school class is randomly assigned to the experimental group, and a different intact class is assigned to the control group. Since there is no randomization of subjects, it is likely that the two groups will differ in many organismic variables. The original difference between $\overline{Y}_E - \overline{Y}_C$ is not interpretable as an estimate of the treatment effect since it is partially confounded with organismic variable differences between the groups. The covariance between A and an organismic variable, x_j, is

$$\delta_{Ax_j} = .5(\mu_{x_jE} - \mu_{x_jC})$$

When the x_j means differ for the two groups, A is no longer orthogonal to x_j.

Evans and Anastasio defend *use two* partly on the basis that "the covariate differences among the groups should be relatively small" (4:228). This is a purely gratuitous assumption. If a school assigned students of a given grade randomly into two classes, the statement should be true, since two different teachers during the present term are unlikely to produce massive behavioral changes. However, if a school uses ability grouping, the differences between the intact groups may be substantial. One suspects *use two* of ANCOVA occurs most often when E's have encountered group differences that are not easy to overlook

If E obtains two organismic variables x_1 and x_2, and uses them in ANCOVA with intact groups, he is again using the model $y = B_1\underset{\sim}{x}_1 + B_2\underset{\sim}{x}_2 + B_a\underset{\sim}{A} + \epsilon$, but with the A vector no longer orthogonal to $\underset{\sim}{x}_1$ or $\underset{\sim}{x}_2$ (The B's used above again represent population values; the reason for different symbols will be apparent later.) Now B_a will be influenced by whether x_1 and x_2 are in the equation. In fact, $B_a = .5\,(\tilde{\mu}_{yE} - \tilde{\mu}_{yC})$ where the $\tilde{\mu}$'s are the "adjusted means" of y after $\underset{\sim}{x}_1$ and $\underset{\sim}{x}_2$ have been "partialled out." It is precisely $2B_a$ that E will attempt to interpret as the "treatment effect."

The dubious nature of this interpretation is shown by the following argument. Of the many possible organismic variables that may be obtained on the subjects, let us assume that only four organismic variables are needed to account for individual variation in y. That is, we shall assume a simple case where all other organismic variables have a $\beta_j = 0$ once these four variables are included in the model. In addition, the treatment has an additive effect so that $\beta_a = .5\,(\mu_{yE} - \mu_{yC})$ is non-zero. Thus, we assume the TRUTH condition is that $y = \beta_1\underset{\sim}{x}_1 + \beta_2\underset{\sim}{x}_2 + \beta_3\underset{\sim}{x}_3 + \beta_4\underset{\sim}{x}_4 + \beta_a\underset{\sim}{A} + \underset{\sim}{e}$.

With usually good insight into the situation, E has correctly identified $\underset{\sim}{x}_1$ and $\underset{\sim}{x}_2$ as pertinent variables, has obtained measures on them, and is using them as covariates. He has not obtained measures on $\underset{\sim}{x}_3$ or $\underset{\sim}{x}_4$. Since $\underset{\sim}{x}_3$ and $\underset{\sim}{x}_4$ are correlated with x_1, x_2, and A, each of these variables may be represented as linear functions of the observed variables

$$\underset{\sim}{x}_3 = c_1\underset{\sim}{x}_1 + c_2\underset{\sim}{x}_2 + c_3\underset{\sim}{A} + \underset{\sim}{e}_3$$

$$\underset{\sim}{x}_4 = d_1\underset{\sim}{x}_1 + d_2\underset{\sim}{x}_2 + d_3\underset{\sim}{A} + \underset{\sim}{e}_4$$

where c_3 is $.5(\tilde{\mu}_{x_3E} - \tilde{\mu}_{x_3C})$, the adjusted mean difference on x_3 when x_1 and x_2 are partialled out, etc. Substituting these equivalences into the TRUTH equation, what is obtained is:

$$\underset{\sim}{y} = \beta_1\underset{\sim}{x}_1 + \beta_2\underset{\sim}{x}_2 + \beta_3(c_1\underset{\sim}{x}_1 + c_2\underset{\sim}{x}_2 + c_3\underset{\sim}{A} + \underset{\sim}{e}_3) + \beta_4(d_1\underset{\sim}{x}_1 + d_2\underset{\sim}{x}_2 + d_3\underset{\sim}{A} + \underset{\sim}{e}_4) + \beta_a\underset{\sim}{A} + \underset{\sim}{e}$$

Rearranging terms to place all multiplicative constants of x_1 together, etc.,

$$\underset{\sim}{y} = (\beta_1 + \beta_3c_1 + \beta_4d_1)\underset{\sim}{x}_1 + (\beta_2 + \beta_3c_2 + \beta_4d_2)\underset{\sim}{x}_2 + (\beta_a + \beta_3c_3 + \beta_4d_3)\underset{\sim}{A} + (\beta_3\underset{\sim}{e}_3 + \beta_4\underset{\sim}{e}_4 + \underset{\sim}{e})$$

It is recalled that the original covariance model using just the x_1 and x_2 variables was written

$$\underset{\sim}{y} = B_1\underset{\sim}{x}_1 + B_2\underset{\sim}{x}_2 + B_a\underset{\sim}{A} + \underset{\sim}{e}$$

It is clear that the B's of the original model contain all the β terms derived from the TRUTH equation. Thus, the B's not only reflect the other organismic variables used as predictors, but *they are also influenced by all other pertinent organismic variables that are NOT used in the analysis.*

Speaking of the analysis of observational data, Tukey (11:118) says, "It is painful to recognize that . . . every measured variable serves more or less as a proxy for all those that are unmeasured, . . ." The present author would prefer to limit the above observation to only organismic variables and nonorthogonal treatment variables. In this example of covariance, note that B_a by no means consists only of the treatment effect β_a. Only if *all* other pertinent unmeasured organismic variables have zero values of c_3 (and d_3, etc.) is the observed covariance "treatment effect" B_a equal to the true treatment effect β_a. That is, for B_a to be equal to β_a, we must believe not only that the analysis accounted for the influence of x_1 and x_2 on y, but also that x_1 and x_2 alone simultaneously remove all possible differences between the two group means in every other pertinent unmeasured x_j variable.

In reality, of course, almost any dependent behavior is influenced by dozens of organismic variables. Thus, there would be dozens of non-zero β_j's and dozens of "differences between the adjusted means of $\underset{\sim}{x}_j$ when the covariance variables have been partialled out" that would have to be zero before $B_a = \beta_a$. It strains credulity to believe that E, with the present state of the arts in education and psychology, can specify and accurately measure two or three covariates that will reduce the c_3, d_3 terms of "all pertinent organismic variables" to zero.

Note the source of this difficulty is that the quasi-experimental designs of *use two* fail to keep the treatment vector orthogonal to all other predictors. In the original three-vector covariance model of *use one*, E also had an incomplete model; $\underset{\sim}{x}_3$ and $\underset{\sim}{x}_4$ of the TRUTH were ignored. However, this in no way changed the value of β_a. Since A is orthogonal to all four organismic variables, it is precisely the same in the four-covariate model as in the two-covariate model as in the ANOVA no-covariate model. The regression coefficients of $\underset{\sim}{x}_1$ and $\underset{\sim}{x}_2$ would be influenced precisely as they are above, so that $\underset{\sim}{x}_1$ and $\underset{\sim}{x}_2$ would serve as proxies for the missing $\underset{\sim}{x}_3$ and $\underset{\sim}{x}_4$, but β_a would be unaffected and the treatment effect correctly assessed.

Note also that the argument has been phrased using parameter values so that no complications arising from sampling fluctuation have besmirched the picture. In addition, it has not been necessary to raise the problem that the $\underset{\sim}{x}_j$'s we use contain measurement error, while the MR models assumed fixed x's. Nor has it been necessary to invoke violations of assumptions (3). Operating under ideal conditions, *use two* of ANCOVA still produces confounded results whenever E has ignored *any* pertinent organismic variables whose means differ from group to group after the covariate measures used have been partialled out. In contrast, in *use one*, ignoring heterogeneous regression slopes is comparable to ignoring a treatment by levels interaction. This would inflate SS_E, and produce a conservative test (compared to that where the true model were known). However, the test still may be more powerful than the ANOVA alternative.

Thus, the ANCOVA is a valuable, robust tool for improving the power of experimental designs where subjects are randomly assigned to treatments. It is *not* a miracle worker that can produce interpretable results from the quasi-experimental designs of *use two.*

REFERENCES

1. Cronbach, L.J.; and Furby, L., "How Should We Measure 'Change' or Should We?," *Psychological Bulletin*, 74: 68-80, 1970.
2. Darlington, R. B., "Multiple Regression in Psychological Research and Practice," *Psychological Bulletin*, 69: 161-182, 1968.
3. Elashoff, J.D., "Analysis of Covariance: A Delicate Instrument," *American Educational Research Journal*, 6: 383-402, 1969.
4. Evans, S.H.; and Anastasio, E.J., "Misuse of Analysis of Covariance When Treatment.Effect and Covariate are Confounded," *Psychological Bulletin*, 69: 225-234, 1968.
5. Ferguson, G.A., *Statistical Analysis in Psychology and Education* (2nd ed.), McGraw-Hill, New York, 1966.
6. Kerlinger, F.N.; and Pedhazur, E.J., *Multiple Regression in Behavioral Research*, Holt, Rinehart & Winston, New York, 1973.
7. Li, J.R., *Statistical Inference II: The Multiple Regression and Its Ramifications*, Edwards Bros., Ann Arbor, Mich., 1964.
8. Lord, F.M., "A Paradox in the Interpretation of Group Comparisons," *Psychological Bulletin*, 68: 304-305, 1967.
9. Lord, F.M., "Statistical Adjustments When Comparing Preexisting Groups," *Psychological Bulletin*, 72:336-337, 1969.
10. McNemar, Q., *Psychological Statistics* (4th ed.), Wiley, New York, 1969.
11. Tukey, J.W., "The Zig-zagging Climb from Initial Observation to Successful Improvement: Comments on the Analysis of National Assessment Data," in W.E. Coffman (ed.), *Frontiers of Educational Measurement and Information Systems—1973*, Houghton Mifflin, Boston, 1973, pp. 113-120.

PERFORMANCE UNDER TRADITIONAL AND MASTERY ASSESSMENT PROCEDURES IN RELATION TO STUDENTS' LOCUS OF CONTROL: A POSSIBLE ATTITUDE BY TREATMENT INTERACTION[1]

CARL H. REYNOLDS
Boston College

J. RONALD GENTILE
State University of New York at Buffalo

ABSTRACT

Previous research in locus of control suggested the hypothesis that internal subjects should perform better under mastery than under traditional assessment procedures, while the reverse should be true of externals. Two experiments were conducted using undergraduate and graduate subjects. Neither the LC nor the assessment procedure main effects were significant in either study, and no interaction was found with the undergraduates. With graduate subjects there was a significant interaction opposite in direction to expectations. Subjects overwhelmingly preferred the mastery procedures. These results are harmful to the construct validity of the I-E Scale (9) and supportive of the mastery learning approach.

ONE OF THE MOST SALIENT differences between mastery learning (1) and traditional educational practice is the amount of control exercised by the student over the educational process. Under a mastery approach the student can usually study at his own pace, decide when he is ready to test his mastery of the material, and determine to a large extent his own course grade. In contrast, under a traditional approach the student must perform more at the instructor's rate and may have less control over his course grade, especially if norm-referenced assessment is being used. The authors were interested in studying this situational difference in the student's control over events important to him as it interacted with the personality construct of locus of control (LC). LC is conceived as a generalized expectancy regarding the control of one's reinforcements (7). A person with an internal LC feels, in general, that he himself is in control of the delivery of his own rewards and punishments. A person with an external LC believes that his reinforcements are regulated by external forces such as luck, powerful others, fate, etc.

Seeman and Evans (11) and Seeman (12) found that internals were more likely than externals to seek out information relevant to their needs. Lefcourt, Lewis, and Silverman (5), Rotter and Mulry (8), and Schneider (10) all reported finding that internals preferred, or took more seriously, situations in which they perceived themselves to be in control, and Watson and Baumal (16) found that internals made fewer errors in a perceived skill than in a perceived chance situation. The reverse findings were true of externals in each of these studies.

In light of the above evidence, the authors hypothesized (*a*) that internal *S*s would prefer an assessment system based on mastery learning to a traditional assessment approach, while the reverse would be true of externals; and (*b*) that internal *S*s would perform better in a mastery learning than in a traditional assessment format, while the reverse would be true of externals. Thus, these research hypotheses provided a test of an aptitude by treatment interaction (2, 3).

Method

Two similar experiments were conducted to test the interaction hypotheses. *Experiment I* involved 76 undergraduate student teachers enrolled in a required course in

educational psychology, and *Experiment II* involved 44 graduate students in a similar graduate level course. Both courses were designed and supervised by the second author, and both were divided by content into four consecutive segments: classroom applications of reinforcement principles; the psychology of discipline; the relationships of beliefs and attitudes to behavior; and measurement and mastery learning theory. Examinations for each unit were scheduled at fixed times, and all students took the same form of the test at that time. For students in the traditional format, the score on that test constituted the basis for a letter grade on that unit. Students in the mastery format had to demonstrate competence in the unit, defined as achieving a score of 80% or more. If the student did not demonstrate competence, he was apprised of his areas of weakness by the instructor or a course assistant and helped to learn the material. When the student felt prepared to demonstrate his mastery of the material, he was given an alternate form of the same test. This process continued until the student achieved mastery.

Experiment II also included a third assessment condition, termed modified mastery, wherein *Ss* who failed initially to attain mastery of the unit were given the option of not restudying the material and not taking another mastery test. Such *Ss* could simply accept a C, say, rather than learn the material to the specified criterion. In this condition, then, students had even more control over the conduct of the course than in the mastery condition.

In Experiment I, *Ss* were assigned to take two segments under the traditional course format and two under the mastery learning format. In *Experiment II* students were assigned to take one of the first three units of instruction under the traditional course format, one under the mastery learning course format, and one under the modified mastery course format. *Ss* were allowed to choose the format they preferred for the last unit. *Experiment II* analyses were based only on the first three units of instruction, since the *Ss* were randomly assigned to conditions for those units only.

All students were pre- and post-tested on an instrument which covered all four units of instruction, and which included a number of items assessing attitudes toward the subject matter and teaching. The I-E Scale (9) was administered during the pre-test to measure LC. Each student's standard score on the section of the post-test corresponding to the unit he took under each assessment condition was employed as the dependent variable. LC was a between-subjects factors, while assessment condition was a within-subjects factor. The analyses were performed in accordance with procedures outlined by Finn (4) and elaborated by Peng (6) for designs which employ correlated groups.

Results

Both the I-E Scale and the post-test instrument showed adequate reliability in both experiments (I-E Scale: $I = .79$, $II = .81$; post-test: $I = .59$, $II = .74$). The students, regardless of LC group, showed an overwhelming verbal preference for the mastery assessment procedures ($I = 68\%$, $II = 70\%$) over either the modified mastery ($I = 26\%$, $II = 30\%$) or the traditional (I & $II = 0\%$) procedures. Since the *Experiment I Ss* did not themselves experience the modified mastery procedure, it was presented as a hypothetical alternative. In *Experiment II,* we had a strong behavioral measure of assessment procedure preference, since the students were allowed to choose the format they preferred for the last unit. Twenty-five (57%) chose the mastery procedures, eighteen (41%) chose the modified mastery conditions, and one (2%) chose the traditional assessment procedure. It is believed this decisive preference for the mastery approach should carry some weight with course planners.

In *Experiment I,* the scores on each subtest of the pre- and post-tests were standardized, and each student was assigned a pre-test mastery score, a pre-test traditional score, a post-test mastery score, and a post-test traditional score by combining his standard scores on the two subtests of each instrument corresponding to the units of instruction taken under mastery or traditional course format. The I-E Scale scores were trichotomized so that scores of 9 or less indicated internality ($N = 25$), scores between 10 and 14 indicated neither internality nor externality ($N = 28$), and scores of 15 or more indicated externality ($N = 23$).

To test for the interaction of LC and course format, the difference was calculated between each *S*'s post-test mastery score and his post-test traditional score. A similar difference score was also calculated for each *S*'s pre-test scores. A one-way analysis of covariance was conducted over the three levels of LC on the post-test difference between mastery and traditional conditions, using the comparable pre-test difference score as a covariate. The results indicated that the pre-test difference (covariate) was not significantly related to the post-test difference ($F < 1$). This was expected, since there was no reason to believe *Ss*' pre-test difference scores should be in any way related to post-test difference scores. The interaction of LC and course format also yielded $F < 1$, which did not support the research hypothesis of this study.

To test for a main effect of mastery versus traditional course format, the mean of the post-test difference scores for all 76 *Ss* was tested to see if it was significantly different from zero. This test yielded $F = 3.3$; $df = 1,72$; $p = .07$. This nearly significant result may have occurred because *Ss* had more experience with the material in the mastery

format units, since they often took a number of tests on those units.

To test for a main effect of LC, the post-test mastery score was combined with the post-test traditional score, and a one-way analysis of covariance was conducted on post-test scores, using the comparable sum of pre-test scores as a covariate. In this case, the covariate was significantly related to the criterion scores. Pre-test scores accounted for 20% of the variance in post-test scores ($r = .45$; $F = 17.9$; $df = 1, 72$; $p < .01$). However, LC was not significantly related to adjusted scores on the post-test ($F < 1$). Tables 1 and 2 summarize these results.

In *Experiment II*, the scores on the first three subtests were again standardized within subtests and across all *S*s, and each *S* received a traditional, a modified mastery, and a mastery score corresponding to his standard scores for the approprate instructional units. This was done for both pre- and post-tests. Since there were fewer *S*s in this experiment, they were simply dichotomized on the I-E Scale (rather than trichotomized as before) into internals with I-E scores of 11 or less ($N = 23$), and externals with scores of 12 or more ($N = 21$).

The logic of the analysis was exactly the same for this experiment as for the earlier one. However, since in this experiment there were two degrees of freedom for the course format factor, two difference scores (mastery versus traditional, and modified mastery versus traditional) were used simultaneously as a multivariate set of dependent variables in order to test for a LC × format interaction and to test for a course format main effect. The test for a LC main effect was again a univariate test employing the sum of scores under all experimental conditions as the dependent variable. Tables 3 and 4 summarize these results.

Again, the pre-test differences between scores for those units taken under mastery conditions and scores for units taken under traditional conditions were found to be unrelated to the same post-test differences. Hence, it was unnecessary to employ, as the authors did, such pre-test difference scores as covariates. The multivariate test of the

Table 1.—*Experiment I* Cell Means and Standard Deviations

Source	Mastery		Traditional	
	Pre	Post	Pre	Post
Internals ($N = 25$)				
$\bar{X}$	−.56	−.16	.10	−.48
SD	1.45	1.23	1.81	1.31
Moderates ($N = 28$)				
$\bar{X}$	.26	.32	.03	.01
SD	1.60	1.42	1.43	1.55
Externals ($N = 23$)				
$\bar{X}$	.17	.43	.34	.04
SD	1.16	1.26	1.46	1.54

Table 2.—*Experiment I* Analyses of Covariance

Source*	*df*	*MS*	*F*	*p*<	Effect tested
Constant term	1	9.59	3.28	.07	Mastery vs. traditional main effect
Between groups	2	.11	.04	.96	Treatment × LC interaction
Covariate $r = .11$	1	2.50	.85	.36	
Error	72	2.93			

*Dependent variable is the difference between mastery and traditional scores on the post-test.
Covariate is the difference between mastery and traditional scores on the pre-test.

Source**	*df*	*MS*	*F*	*p*<	Effect tested
Between groups	2	3.86	.92	.40	Locus of control main effect
Covariate $r = .45$	1	75.18	17.93	.001	
Error	72	4.19			

**Dependent variable is the sum of the mastery and traditional scores on the post-test.
Covariate is the sum of mastery and traditional scores on the pre-test.

LC × course format interaction was marginally significant ($F = 3.0$; $df = 2,39$; $p < .06$), and the univariate tests on each of the post-test difference scores, using the appropriate pre-test difference scores as covariates, were clearly significant ($Fs = 5.2$ and 6.1; $df = 1, 41$; $ps < .04$ and $.02$). However, as shown in Figure 1 (the means graphed in Figure 1 are the uncorrected means, since the covariates were not effective in the test of the interaction), this significant interaction was opposite in direction to the hypothesis! The stronger the external control of the course, the better the internals did. Contrarily, the greater the opportunity for self-direction, the better the externals performed. Nothing in the theory of LC would suggest that this should be the case. In *Experiment I* no interaction was found, and in *Experiment II* an interaction opposite in direction to the hypothesis was found. These contradictory results suggest a need for replication, but both experiments agreed in failing to confirm the hypothesis. In neither experiment were there any significant differences between LC groups with respect to preference for assessment procedures. Overwhelming preference for the mastery approach was the rule regardless of I-E Scale score.

The cognitive impact of the course was demonstrated by highly significant ($p < .01$) changes in performance

Table 3.—*Experiment II* **Cell Means and Standard Deviations**

	Modified mastery		Mastery		Traditional	
	Pre	Post	Pre	Post	Pre	Post
Internals ($N = 23$)						
$\bar{X}$	.00	−.20	−.07	−.05	.38	.41
SD	.96	.85	.91	.67	.95	.78
Externals ($N = 21$)						
$\bar{X}$	−.07	.09	−.22	.04	−.06	−.30
SD	1.03	1.24	1.12	.93	.92	1.26

Table 4.—*Experiment II* **Analyses of Covariance**

Source*	Univariate *df*	*MS*	*F*	*p*<	Multivariate *df*	*F*	*p*<	Effect tested
Constant term:								
M-T	1	.001	.001	.97				
MM-T	1	.26	.17	.69	2,39	.18	.84	Treatment main effect
Between groups:								
M-T	1	5.27	4.66	.04				
MM-T	1	9.24	5.85	.02	2,39	3.04	.06	Treatment × LC interaction
Covars:								
M-T Mult. $r = .32$	2	2.64	2.33	.11				
MM-T Mult. $r = .21$	2	1.46	.92	.41	4,78	1.20	.32	
Error:								
M-T	40	1.13						
MM-T	40	1.58						

*Dependent variables are (*a*) difference between mastery and traditional scores on the post-test: M-T; (*b*) difference between modified mastery and traditional scores on the post-test: MM-T.
Covariates are (*a*) difference between mastery and traditional scores on the pre-test; (*b*) difference between modified mastery and traditional scores on the pre-test.

Source**	*df*	*MS*	*F*	*p*<	Effect tested
Between groups	1	.06	.012	.91	Locus of control main effect
Covariate $r = .32$	1	22.61	4.59	.04	
Error	41	4.93			

**Dependent variable is the sum of post-test scores under all three conditions.
Covariate is the sum of pre-test scores under all three conditions.

(2.5 to 3.0 pre-test standard deviations) from pre- to post-test. The affective impact of the course was apparent in a significant ($p < .01$) positive shift–sign test (13)–in the attitudes of these *Ss* toward the concepts and principles of the course and their applications to teaching.

Thus, while the course had powerful cognitive and affective effects, neither LC nor assessment condition had a significant effect, and the hypothesized interaction failed to appear.

Discussion

Several explanations may be advanced to account for the data. Originally, those who promulgated the LC construct hypothesized that it would be strongly related to *n*-achievement (7), which, one would expect, would lead to school achievement. Perhaps, however, LC simply is not a powerful variable in school situations. Rotter (9) and Warehime (15) have suggested as much in efforts to account for the fact that the I-E Scale seems to be unrelated to school grade point average. The hypothesis of a relationship between LC and *n*-achievement has also fared poorly. Wolk and DuCette (17) found no significant correlation between the I-E Scale and two measures of *n*-achievement in two samples of *Ss*.

Another possible explanation of the findings is that the I-E Scale assesses socio-political attitudes rather than an underlying personality dimension with motivational consequences. The responses to the I-E Scale which indicate an internal LC usually emphasize individualism and success through hard work. Such responses should be congenial to those of conservative socio-political philosophy. On the other hand, external responses often emphasize collectivism and common oppression by greater powers. These responses probably fit well in the world-view of many liberal thinkers. Indeed, Thomas (14) found that although his sample of 30 liberals was more politically active than his sample of 30 conservatives, the liberals were significantly more external than the conservatives.

If the I-E Scale measures socio-political philosophy, the interaction found in the second experiment is readily explained. If the externals are liberals, they should prefer the more liberal course formats, while the conservative internals should prefer the traditional instructional methods. This is exactly what was found in *Experiment II.*

However interpreted, the results of these experiments are damaging to the construct validity of the I-E Scale. Further experimentation should be undertaken to resolve the discrepancies between the results of the two studies, but there is no evidence in either experiment of the interaction predicted by LC theory.

The finding of most importance for education was that both undergraduate and graduate students showed an overwhelming preference for the mastery learning format. Since the students learned the material equally well under all of the assessment procedures, the authors believe this result argues strongly in favor of the mastery learning approach.

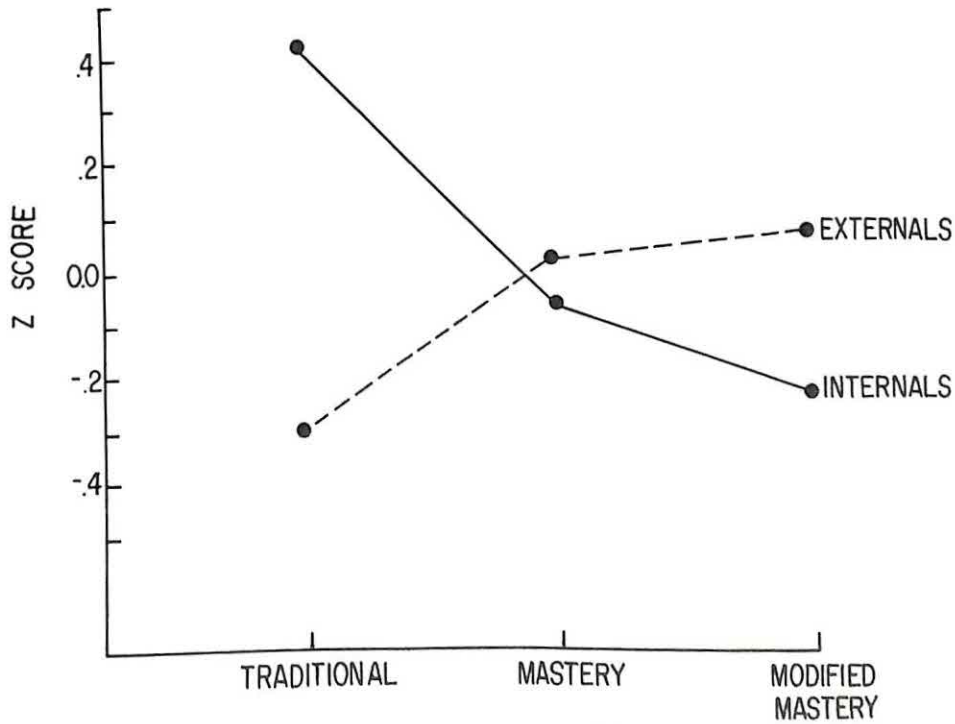


Figure 1.–Post-test performances of internals and externals in the different assessment conditions

NOTE

1. Appreciation is expressed to Dr. Howard Kight and Mr. Bruce Kestleman for their suggestions; to the students in these courses for their willingness to be part of an experiment; and to the following course assistants for helping to make these courses successful: John Dilendik, Marilyn Dozoretz, Murial Frank, Kathleen Van Every, Marvin Lew, Alfred Sarnowski, Joseph Zampogna, and Margaret Zabranskey. An earlier version of this paper was presented on April 2, 1975, at the American Educational Research Association Annual Meeting in Washington, D. C.

REFERENCES

1. Block, J. H. (ed.), *Mastery Learning: Theory and Practice,* Holt, Rinehart & Winston, New York, 1971.
2. Bracht, G. H., "Experimental Factors Related to Aptitude-Treatment Interactions," *Review of Educational Research,* 40:627-645, 1970.
3. Cronbach, L. J.; and Snow, R. E., "Individual Differences in Learning Ability as a Function of Instructional Variables," Final Report, ERIC Document No. ED 029 001, U. S. Office of Education, March 1969.
4. Finn, J., "Multivariate Analysis of Repeated Measures Data," *Multivariate Behavioral Research,* 4:391-413, 1969.
5. Lefcourt, H. M.; Lewis, L.; and Silverman, I. W., "Internal versus External Control of Reinforcement and Attention in a Decision-making Task," *Journal of Personality,* 36:663-682, 1968.
6. Peng, S., "Analysis of Repeated Measures Data," manuscript prepared for an American Educational Research Association Annual Meeting Training Session, Washington, D. C., April 1975.
7. Rotter, J. B.; Seeman, M.; and Liverant, S., "Internal versus External Control of Reinforcement: A Major Variable in Behavior Theory," in N. Washburne (ed.), *Decisions, Values and Groups,* Pergamon Press, New York, 1962, pp. 473-516.
8. Rotter, J. B.; and Mulry, R. C., "Internal versus External Control of Reinforcement and Decision Time," *Journal of Personality and Social Psychology,* 2:598-604, 1965.
9. Rotter, J. B., "Generalized Expectancies for Internal versus External Control of Reinforcement," *Psychological Monographs: 80,* No. 609, 1966.
10. Schneider, J. M., "Skill versus Chance Activity Preference and Locus of Control," *Journal of Consulting and Clinical Psychology,* 32:333-337, 1968.
11. Seeman, M.; and Evans, J. W., "Alienation and Learning in a Hospital Setting," *American Sociological Review,* 27: 772-782, 1962.
12. Seeman, M., "Alientation and Social Learning in a Reformatory," *American Journal of Sociology,* 69:270-284, 1963.
13. Siegel, S., *Nonparametric Statistics,* McGraw-Hill, New York, 1956.
14. Thomas, L. E., "The I-E Scale, Ideological Bias, and Political Participation," *Journal of Personality,* 38:273-286, 1970.
15. Warehime, R. G., "Generalized Expectancy for Locus of Control and Academic Performance," *Psychological Reports,* 30:314, 1972.
16. Watson, D.; and Baumal, E., "Effects of Locus of Control and Expectation of Future Control upon Present Performance," *Journal of Personality and Social Psychology,* 6:212-215, 1967.
17. Wolk, S.; and DuCette, J., "Locus of Control and Achievement Motivation: Theoretical Overlap and Methodological Divergence," *Psychological Reports,* 29:755-758, 1971.

DIRECTIONS FOR J.E.E. CONTRIBUTORS

The Journal of Experimental Education publishes specialized or technical education studies, treatises about the mathematics or methodology of behavioral research, and monographs of major current research interest.

ABSTRACT REQUIRED

An abstract of not more than 120 words must accompany each manuscript. It should precede the text, be designated *ABSTRACT,* and conform to the following standards:

1. In a research paper, include statements of (a) the problem, (b) the method, (c) the data, and (d) the conclusions.

2. In a review or discussion article, state the topics covered and the central thesis.

3. Standard abbreviations may be used freely if the meaning is clear. Use complete sentences and do not repeat information which is in the title.

TEXT

Usually research articles should have a clearly organized order of presentation. The following elements and their order is a guide and is not intended to be prescriptive.

The Problem. The nature, scope, and significance of the problem should be presented.

Related Research. Presentation and discussion of related research should be selective and should include references essential to clarify the problem under examination.

Methodology. This section should consist of hypotheses, description of the sample and sampling procedures, discussion of the variables, description of the data-gathering instruments (if well-known, their description may be omitted), and general description of the statistical procedures.

Presentation and Analysis of Data. Analysis of the data and conclusions about the hypotheses should be more than mere presentation. Rival hypotheses and significance for related studies and educational theory should be considered. Summary data (means, standard deviations, frequencies, correlation coefficients, etc.) should be presented in tabular or graphic form. Discussion of these data should be interpretive.

Summarizing Statements. A summary of conclusions and implications for education may supplement the abstract.

STYLE

Certain suggestions are offered here to achieve uniformity in publication style and to conserve the time and energy of the author and editorial staff in the preparation of copy. *A Manual of Style,* 12th ed., University of Chicago Press, Chicago, 1960, may be used as a style manual in preparation of manuscripts.

Two Copies Required. Copies should be double spaced with wide margins. Typed copies are preferred, but dittoed or mimeographed copies will be accepted if they are legible.

Subheads. Articles are frequently improved by the judicious use of subheads. However, avoid use of the title, INTRODUCTION, for a lead section.

Title. Try to use a short title, preferably no more than ten words. Avoid superfluous phrases, such as "A Comparison of . . .," "A Study of . . .," and "The Effectiveness of . . . ."

Tables. Prepare tables precisely as they are to appear in *The Journal,* caption each with a brief and to-the-point title, and number consecutively with arabic numerals: *Table 3.* INTERCORRELATION MATRIX, VARIABLES OPTIMALLY ORDERED.

Figures. Draw any graphs and charts in black ink on good quality paper, title each, and number consecutively, thus: *Figure 4.* SCHOOL ENROLLMENT. Send the original copy. We cannot accept mimeographed figures or charts. Data should not be framed, and ordinates and abscissas should be shortened to occupy as little space as possible.

Tables and Figures. Tables and figures must be original copies acceptable for reproduction. A charge will be assessed for any redrawing or re-typing of tables or figures.

Technical Symbols. All symbols, equations, and formulas must be clearly represented. Differentiate between numbers and letters (zero and the letter o; one and the letter l, etc.) Identify hand-drawn Greek letters in the margin. Align subscripts carefully.

Footnotes. Avoid explanatory footnotes by incorporating their content in the text. However, for essential footnotes, identify them with consecutive superscripts, as *book,*[2] *study,*[3] etc., and list the footnotes in a section, entitled FOOTNOTES, at the end of the text, but preceding the REFERENCES.

References. References should be listed alphabetically according to the author's last name at the end of the manuscript. Each entry should be numbered. Citation of a reference in the text should be made with this number, as (2); or if specific pages are to be indicated, as (2:245–7).

Illustrations of article and book references:

1. Bittner, Reign H. and Wilder, Carlton E., "Expectancy Tables: A Method of Interpreting Correlation Coefficients," *The Journal of Experimental Education,* 14:245–52, March 1946.
2. Thomson, Godfrey, H., *The Factorial Analysis of Human Ability,* Houghton Mifflin Co., Boston, 1950, 383 pp.

PROCEDURES

Send manuscripts to John Schmid, Department of Research and Statistical Methodology, University of Northern Colorado, Greeley, CO 80631.

Each contributor will receive 2 complimentary copies of the issue in which his article appears. Keep an exact carbon copy of your manuscript so that if a question arises the editors can refer you to specific pages, paragraphs, or lines for clarification by letter.

THE JOURNAL OF EXPERIMENTAL EDUCATION

4000 Albemarle Street, N.W., Suite 302,
Washington, D.C. 20016

Return Postage Guaranteed

Second Class
Postage Paid at
Washington, D.C.

THE *Journal* OF Experimental Education

Volume 44, Number 1 Fall 1975

In this issue:

THE *Journal* OF EXPERIMENTAL EDUCATION

Volume 44, Number 1 Fall 1975

CONTENTS

The Journal of Experimental Education is published four times a year by HELDREF publications, 4000 Albemarle St., N.W., Washington, D.C. 20016. Annual subscription rates are $12.50 for institutions and $10 for individuals, plus $3 postage for all subscriptions outside the United States and Canada. Single copies $3. Second class postage paid at Washington, D.C. All business correspondence should be sent to this address. Claims concerning missing issues made within 6 months will be serviced free of charge. Send all manuscripts to Prof. John Schmid, Department of Research and Statistical Methodology, University of Northern Colorado, Greeley, CO 80631.

Arvil S. Barr, Founder

EDITOR AND PUBLISHER • 1932-1962

(The Journal of Experimental Education is indexed/abstracted in Abstr S.W., CSPA, Current Contents, Ed Adm Abst, Educ Ind, Soc. of Ed. Abst., Current Index to Journals in Education, Language and Language Behavior Abst.)

ATTITUDE-APTITUDE RELATIONSHIPS IN THE QUANTITATIVE DOMAIN: A CANONICAL ANALYSIS

ANDREW G. BEAN
CATHLEEN KUBINIEC MAYERBERG
Temple University

ABSTRACT

The purpose of this study was to investigate the relationships among nine measures of attitude toward quantitative concepts and verbal and quantitative aptitude. Subjects were 353 graduate students enrolled in educational research courses. The attitude measures consisted of nine factor scores drawn from a semantic differential instrument; the aptitude measures were the Graduate Record Examination Verbal and Quantitative scores. Canonical correlation was used to relate the attitude to the aptitude measures. Two significant canonical variates were obtained. The first showed a moderate relationship between positive attitudes toward quantitative concepts and high quantitative aptitude. The second indicated a slight relationship between negative attitudes toward quantitative concepts and high verbal aptitude.

MOST MEASURES OF "attitude toward mathematics" consist of a single scale based on items which sample a variety of attitudes toward different aspects of mathematics rather than focusing on some specific part of the subject. Such generalized instruments may fail to measure important facets of the variable of interest (2). In support of this idea, Mayerberg and Bean (5) reported data indicating that attitude toward different mathematics-related concepts (e. g., Algebra, Statistics, Calculations, Formulas, etc.) could be considered as multidimensional. They suggested use of the more general term "attitude toward quantitative concepts" instead of "attitude toward mathematics" to describe this domain.

Since aptitude is related to achievement, and since achievement affects attitude and vice versa, one would expect attitude to be related to aptitude. Aiken (2) cites several studies which report low to moderate correlations between attitudes toward mathematics and measures of scholastic aptitude.

In all such studies, however, attitude was measured by a single scale. Needed, then, is a study which examines the relationship between the various aspects of attitude toward quantitative concepts and scholastic aptitude. The purpose of this study was to investigate the relationships among nine relatively independent factors of attitude toward quantitative concepts and measures of quantitative and verbal aptitudes.

Studies investigating attitude-aptitude relationships in the quantitative domain support the generalization that attitude toward mathematics is more closely related to quantitative aptitude than to verbal aptitude. For example, in a sample of 40 college students, Dreger and Aiken (3) found that anxiety toward mathematics had a correlation of −.25 with the American Council on Education (ACE) quantitative score and −.08 with ACE linguistic score. Aiken (1) also found a correlation of .37 between Mathematics Attitude Scale (MAS) score and Scholastic Aptitude Test (SAT) quantitative score, but no significant correlation with SAT verbal score.

The relationship between mathematics self-concept and various aptitudinal variables has also been studied. Using a sample of seventh-grade students, Holly *et al.* (4) reported a significant correlation of .47 between scores on the Mathematics Self-Concept Scale (MSCS) and the Comprehensive Test of Basic Skills Pretest in Mathematics. Correlations of similar magnitude were obtained between the MSCS and

other measures of mathematics aptitude, compared with a correlation of .32 with the Lorge-Thorndike Intelligence Test Verbal I. Q. Thus, from two related perspectives—attitude toward the subject matter and perception of one's ability in the subject matter—there is evidence of a relationship between aptitude in and attitude toward mathematics.

From the above studies, one would expect measures of attitude toward quantitative concepts to be moderately related to quantitative aptitude; furthermore, they should be more highly related to quantitative aptitude than to verbal aptitude. Thus, it is hypothesized that a canonical correlation analysis relating several relatively independent measures of attitude toward quantitative concepts to measures of verbal and quantitative aptitude will result in a single significant canonical variate. All of the attitude measures as well as the quantitative aptitude measure will be highly correlated with this first variate. That is, it is predicted that only one variate is necessary to explain the inter-relationships between the attitude and aptitude domains. The nature of this variate will be related to quantitative aptitude, but not to verbal aptitude.

Method

Subjects consisted of students enrolled in one of three graduate-level courses in educational research and statistics offered within the College of Education of a large urban university. To obtain an adequate sample size, data were gathered from all students enrolled in these courses for four different semesters. The total sample size was 353 (175 males and 178 females).

Variables

The nine attitude measures used were obtained by having subjects respond to a semantic differential measuring instrument containing six quantitative concepts and fourteen bipolar adjective scales. Three of the concepts (Algebra, Statistics, and Mathematics) reflected substantive areas in the quantitative domain; the remaining three concepts (Numbers, Calculations, and Formulas) reflected tools employed in that quantitative domain.

The six concepts were rated on the following fourteen bipolar adjective scales: (1) enjoyable-unenjoyable; (2) attractive-repellent; (3) interesting-boring; (4) pleasant-unpleasant; (5) valuable-worthless; (6) useful-useless; (7) important-unimportant; (8) simple-complex; (9) easy-difficult; (10) lucid-obscure; (11) clear-hazy; (12) meaningful-meaningless; (13) intelligible-unintelligible; and (14) good-bad. These adjectives were selected on the basis of two criteria: (1) a meaningful relationship to the concepts rated, and (2) a presumably heavy loading on the evaluative meaning dimension.

From a principal factor analysis with oblique rotation for simple loadings, nine factors were obtained and labeled as follows: (1) Algebra; (2) Statistics; (3) Calculations; (4) Formulas; (5) Numbers and Mathematics; (6) Useful; (7) Easy; (8) Clear; and (9) Good. The first five factors reflect attitude toward a specific quantitative concept. For example, a person with a high score on the first factor could be described as having a positive attitude toward Algebra, describing it as enjoyable, interesting, etc. The remaining four factors represent a generalized attitude toward the entire domain of quantitative concepts. For example, a person with a high score on the sixth factor describes all of the quantitative concepts as "Useful."

Correlations among the nine factors were generally low and positive. A thorough description of these factors, along with evidence of their reliability and construct validity, is given in Mayerberg and Bean (5).

Factor scores for each of the nine factors described above served as the attitude measures used in the study. The measures of scholastic aptitude consisted of Graduate Record Examination Verbal (GREV) and Quantitative (GREQ) scores.

Data Analysis

Canonical correlation analysis was used to relate the predictor set of attitude measures to the criterion set of aptitude measures. To aid in the interpretation of the canonical variates, the variable-variate correlation matrix was computed, providing canonical component loadings (6).

Results

Means, standard deviations, and correlations among the predictors and the criteria were computed separately for each sex. Since the correlations for males and females were similar, all of the data presented here are based on the sexes combined.

The attitude factors were in z-score form; thus, all attitude measures had a mean of zero and a standard deviation of one. GREV and GREQ means were 547 and 515 respectively; the standard deviations were 104 and 102, respectively.

Table 1 presents the correlations among the predictors and the criteria. As stated previously, correlations among the attitude measures were generally low and positive. All correlations of the attitude measures with GREQ were positive and statistically significant at the .05 level. Three factors, "Algebra," "Easy," and "Clear" showed correlations of from .35 to .39. Thus, positive attitudes toward quantitative concepts were associated with high quantitative aptitude.

The magnitude of correlations between the attitude measures and GREV were quite low, with the only statistically significant correlations being negative. Four factors, "Numbers and Mathematics," "Calculations," "Easy," and "Good" showed marginally significant correlations ranging from $-.17$ to $-.11$.

Table 1.— Intercorrelations among Attitude and Aptitude Measures

	2	3	4	5	6	7	8	9	10	11
Attitude Measures										
1. Algebra	.32	.21	.43	.30	.33	.27	.31	.37	.00	.39
2. Statistics		.26	.37	.08	.20	.26	.24	.27	-.05	.23
3. Calculations			.32	.17	.17	.30	.24	.24	-.11	.16
4. Formulas				.14	.25	.33	.31	.30	-.01	.27
5. Numbers & Math					.07	.22	.26	.23	-.17	.14
6. Useful						-.12	.25	.45	.07	.20
7. Easy							.40	.12	-.11	.35
8. Clear								.33	.07	.37
9. Good									-.12	.15
Aptitude Measures										
10. GREV										.30
11. GREQ										

Note: For N=353, a correlation of .11 or greater is statistically significant at the .05 level, using a two-tailed test.

Canonical correlation analysis of the relationship between the attitude factors and the aptitude measures yielded two significant canonical variates. The canonical correlations of .55 and .30 both were significant at the .01 level.

The two canonical variates may be interpreted by examining the canonical component loadings shown in Table 2. All attitude measures have positive correlations exceeding .30 with the first variate. GREQ loads .93 on the first variate, while GREV shows a near-zero relationship.

Thus, the first variate is consistent with the expectation that positive attitudes toward quantitative concepts are associated with high quantitative aptitude. Attitude measures contributing most to the first variate (specifically those with loadings above .50) are positive attitudes toward "Algebra" and "Formulas" and the attitudes that quantitative concepts are "Easy" and "Clear." The first canonical correlation of .55 indicates a shared variance between the predictor and criterion set of approximately 30 percent.

Contrary to the original hypothesis, a second statistically significant vector was found. Three attitude factors show negative loadings with absolute values exceeding .30. "Numbers and Mathematics" shows the closest relationship to the

Table 2.–Canonical Component Loadings for Two Canonical Variates

	I	II
Predictor Variables		
Algebra	.73	.11
Statistics	.47	-.11
Calculations	.39	-.31
Formulas	.53	.02
Numbers & Math	.39	-.53
Useful	.35	.28
Easy	.75	-.29
Clear	.67	.30
Good	.37	-.34
Criterion Variables		
GREV	-.06	.99
GREQ	.93	.36
Canonical R	.55**	.30**

** $p < .01$

second variate (–. 53). Thus, *negative* attitude toward "Numbers and Mathematics" is associated with higher GREV scores. The second canonical correlation accounts for approximately 9 percent of the shared variance between the predictor and criterion set.

Discussion

Canonical analysis relating nine measures of attitude toward quantitative concepts to verbal and quantitative aptitude resulted in two significant canonical variates. In agreement with the original hypothesis, the first canonical variate related positive attitudes toward quantitative concepts primarily to high quantitative aptitude. It is reasonable to expect that high loadings on this first variate would occur for "Algebra" and "Formulas," since these skills are important in obtaining a high GREQ score. Similarly, persons expressing the attitude that quantitative concepts are "Easy" and "Clear" are likely to say so because of previous successes in quantitative performance.

The judgment that quantitative concepts are "Useful" is not strongly related to GREQ. Such a result is not surprising

since attitudes about the usefulness of quantitative concepts are not strongly related to the perceived ease and clarity of these same concepts (see Table 1). The relatively low correlation between attitude toward "Number and Mathematics" and GREQ is somewhat unexpected. More will be said about this below.

The second canonical variate was essentially a negative relationship between "Numbers and Mathematics" and verbal aptitude. A theoretical explanation for this finding is not easily constructed. One possible empirical explanation for this finding could be that the "Numbers and Mathematics" factor lacks construct validity. Results of the factor analysis used to create the nine attitude factor scores indicated that this factor was the least "clean" in terms of factor structure. It could be that the concept "Numbers and Mathematics" embedded in an instrument containing more specific concepts such as "Algebra" and "Statistics" is too broad and ambigious to provide useful attitude measures.

In future studies, these attitude measures can be examined as predictors of achievement in courses dealing with quantitative topics. The moderate degree of relationship between attitude and aptitude allows for the possibility that attitude can be used to increase predictive validity above that obtained using aptitude measures alone. If the relationship between attitude and aptitude were quite high, then little predictive information would be gained by adding attitude measures to a predictor set already containing a measure of quantitative aptitude.

In summary, positive attitudes toward quantitative concepts were found to be moderately related to quantitative aptitude. Positive attitudes toward "Algebra" and "Formulas" and the attitudes that quantitative concepts are "Easy" and "Clear" are the four measures most closely related to high quantitative aptitude. Other attitude measures, notably a positive attitude toward "Numbers and Mathematics," have a small but statistically significant relationship to lower verbal aptitude.

REFERENCES

1. Aiken, Lewis R., "Personality Correlates of Attitude toward Mathematics," *Journal of Educational Research,* 56:476-480, 1963.
2. Aiken, Lewis R., "Attitudes toward Mathematics," *Review of Educational Research,* 40: 551-596, 1970.
3. Dreger, Ralph M.; and Aiken, Lewis R., "The Identification of Number Anxiety in a College Population," *Journal of Educational Psychology,* 48: 344-351, 1957.
4. Holly, Keith A.; Purl, Mabel C.; Dawson, Judith A.; and Michael, William B., "The Relationship of an Experimental Form of the Mathematics Self-Concept Scale to Cognitive and Noncognitive Variables in a Sample of Seventh-Grade Pupils in a Middle-Class Southern California Community," *Educational and Psychological Measurement,* 33: 505-508, 1973.
5. Mayerberg, Cathleen K.; and Bean, Andrew G., "The Structure of Attitude toward Quantitative Concepts," *Multivariate Behavioral Research,* 9: 311-324, 1974.
6. Weiss, David, "Canonical Correlation Analysis-Counseling Psychology Research," *Journal of Counseling Psychology,* 17: 477-483, 1970.

PERFORMANCE OF READABILITY FORMULAS UNDER CONDITIONS OF RESTRICTED ABILITY LEVEL AND RESTRICTED DIFFICULTY OF MATERIALS[1,2,3]

NELSON RODRIGUEZ T.
Caracas, Venezuela

LEE H. HANSEN
Madison, Wisconsin Public Schools

ABSTRACT

Readability formulas are designed to provide quantitative estimates of the relative difficulty of pieces of writing. Formulas developed to date have been designed to be used across materials of varying difficulty and with subjects of varying ability and maturity. This study explored the extent to which an increase in the accuracy of a specific readability formula could be obtained by norming it for a restricted set of reading materials and subjects. Using the earlier work of Bormuth, the authors constructed readability formulas from cloze data gathered on textbook, newspaper, and leisure reading passages administered to seventh graders. Generally, it was found that narrow-band formulas designed for material of restricted difficulty and subjects with a narrower range of ability offered somewhat more accuracy than more general formulas.

READABILITY FORMULAS are predictive devices that provide quantitative estimates of the relative difficulty of pieces of writing. The general purpose for their use is to estimate the probable success a reader will have in understanding a set of materials without requiring the reader to take tests of any kind (4).

The underlying assumption of readability formulas is that the difficulty of a piece of writing is determined, at least partially, by elements contained in the writing itself: content, style, print, etc. Most readability formulas use style elements for their predictions. The general procedure in the development of readability formulas is the following:

1. A set of plausible readability variables is generated on the basis of previous experience or of linguistic research.
2. Those variables are computed for a set of passages.
3. For a sample of individuals from the target population, scores are obtained in a criterion of comprehension, usually a test based on the passages.
4. Finally, the formula is computed by regressing the readability variables on the criterion score.

Once the formula is obtained, it can be used with subjects of similar characteristics as those in the normalization sample to predict their criterion scores, and, indirectly, their comprehension of the materials.

Klare (4) presented a developmental history of readability formulas up to the 1960s. Although some improvements in computation and cross-validation procedures had previously been reported, a real breakthrough occurred with the development of the cloze procedure, recent advances in linguistic research, and the development of more powerful computer soft- and hardware (1). The cloze procedure allowed for a more objective external criterion and more powerful psychometric techniques; the advances in linguistic research permitted the development of variables that were unknown before; the advances in computer technology allowed for more complex data analysis at a reasonable cost.

Two documents by Bormuth (1, 2) report attempts to incorporate both cloze tests and new linguistic variables into readability research. He reports formulas that reach an unprecedented degree of accuracy, not only in the standardization samples, but also in cross-validation.

The present study was undertaken under the following rationale. In his normative and cross-validation studies, Bormuth used samples of materials of wide-ranging difficulty levels; his samples of students were also from a wide ability range. Under those circumstances, he determined what could be considered a ceiling in the prediction ability of his formulas. It would be interesting to determine a "floor" by restricting the range of the materials as well as the ability of the students. This situation is provided by testing students from one single grade level on materials from one single subject area. Although within one grade level there is usually a wide range of ability, that range will obviously be more restricted than across two or more grade levels. The former is the situation usually confronting the individual teacher when she has to decide what materials to assign for a specific group of students in her class. The teacher would presumably randomly select passages from the books, compute the variables required by the formula, and obtain an estimate of their difficuly. This, together with her knowledge of the students' reading ability (from tests, inventories, and direct observation), would allow her to match students with materials. This is also the situation that school systems confront when deciding what materials to purchase for specific grade levels.

In this context, that is, given a restricted level of ability and a restricted range of difficulty of the materials, the following questions are of interest:

1. How well would Bormuth's formulas predict mean (cloze) difficulty?
2. If a set of variables originally tested by Bormuth were tested under these circumstances, would the same variables enter a multiple regression equation?
3. How well would the set of variables entering the equation predict the (cloze) mean?
4. As contrasted with Bormuth's formulas that were developed under other circumstances, how well would a formula developed in a sample of restricted ability and difficulty ranges predict (cloze) difficulty in other sets of materials?

In this project two studies were performed. The first one attempts to answer the first three questions; that is, (1) to determine how well Bormuth's simple formulas perform in this case; (2) to suggest which variables are the best single set of predictors; and (3) to provide a multiple regression equation. The second study cross-validates Bormuths's formulas and the formula obtained in the first study in two other sets of materials for purposes of comparison.

Procedure

The data used in this project originated from an assessment of reading literacy performed in the Public School System of Madison, Wisconsin. In that project a large number of fourth, seventh, tenth, and twelfth graders were tested using 10-item cloze tests developed on 60- to 70-word passages randomly selected from a predefined universe of materials that the students were supposed to be able to read. That universe included several domains, among others—safety materials, occupational information, textbooks, leisure-time reading, consumer materials, etc. (3). For each passage, the mean cloze score and some additional information was available.

For the present project, the seventh grade data were selected, beginning with textbook, leisure-time materials, and newspapers. There were respectively 45, 36, and 23 passages available in those domains. For each passage a set of linguistic variables was computed. They were used as independent variables to predict mean cloze scores.

All the variables used in this project were developed and used by Bormuth in his 1966 and 1969 readability studies.[3] They were included in this project for one of the following three reasons:

1. They had entered the stepwise multiple regression equation in Bormuth's studies.
2. They did not enter the equations, but showed a high correlation with the criterion.
3. They were included in the manual computation formulas developed by Bormuth.

The second set of variables was included because sampling errors sometime prevent a variable from entering an

Table 1.—Variables Selected for the Present Study from Bormuth's 1966 and 1969 Studies and His Readability Formulas

Variable	1966	1969	Formula
Letter/independent clause	x		
Nouns/structural words	x		
Letter/words	x		
Pronouns/conjunctions	x		
Syllables/sentences		x	x
Syllables/words		x	x
Personal pronouns/words		x	x
Dale-Chall List 769/words		x	x
Dale-Chall List 3000/words		x	x
Words/sentences		x	x
Letters/words		x	x
Letters/meaningful punctuation unit		x	x
Syllables/independent clause	x*		
Words/independent clause	x*		
Structural words/nouns	x*		
Adjectives/structural words	x*		
Anaphora/words		x*	
Letters/syllables		x	
Structural words/adjectives		x*	

* Denotes "promising variables." See text for explanation.

equation. Under the changed circumstances of this project, it was considered that other variables would enter the equations. Table 1 lists the variables and the study in which they originated. Marked with an asterisk are "promising variables," that is, those considered in item number 2 above.

Study I

In this study 45 passages from textbook materials were used. Three statistical analyses were performed:

1. By using Bormuth's formulas, four predicted means were computed for each passage. A zero order correlation was then computed between predicted and obtained mean cloze scores.
2. A stepwise multiple regression equation was computed with nineteen dependent variables (Table 1) and one independent variable (mean cloze). The program was allowed to run forward and unrestricted.
3. The same program (as in 2) was run but restricted at the levels of significance of .05 for inclusion and .25 for exclusion.

Results and Discussion

In considering the results of this study, it must be understood that the present project differs from Bormuth's studies in at least two main points. First, this study uses a more restricted range of passage difficulty as well as ability level of the subjects; second, the average passages used in this study are approximately one-fourth of the length of the ones used by Bormuth (70 words in this study as compared to 287 words in Bormuth's 1966 study and 110 in the 1969 study). Furthermore, Bormuth used five cloze forms for each passage, including, by so doing, each possible word as a cloze item; this project used only a random sample of 20% of the words. The first one is a "built in" difference; the second results from using available data.

How well do Bormuth's formulas perform in this situation?

Table 2 shows the zero order correlation coefficients between predicted and obtained mean cloze scores.

The results show what can be considered a floor value for Bormuth's formulas. Formula 3 seems to be performing better than the others, a result that was confirmed in Study II. All correlations reached significance at the .01 level except Formula 1. When comparing these results with Bormuth's cross-validation studies, a drop in the correlations is observed from around .90 to around .40. This decrease in validity can be due either to the restricted range of difficulty and ability or to the shorter length of the passages. An inspection of the raw data shows that at least one of the variables included in the formulas, personal pronouns per word, was absent in many of the passages. This suggests the likelihood of a lower reliability of the linguistic variables included. Furthermore, in his studies Bormuth found a certain degree of curvilinearity in the scatter plots of expected versus obtained difficulties in his cross-validation studies. It is possible that for shorter passages the relationship between linguistic variables and cloze scores is curvilinear; this would also contribute to reducing the correlation coefficients. This possibility is supported by Bormuth's finding (1) that at the level of independent clause many variables show curvilinearity, whereas at passage level, curvilinearity tended to disappear; this suggests that with increased length in the passages, the relationships are linear.

As a continuation for this study, it is suggested that a study be performed using longer passages (250+ words) but restricting, as in this project, ability range and difficulty. If it is demonstrated that part of the reduction in predictability can be attributed to passage length, it would be necessary to use longer passages in the evaluation of materials.

Would the same variables enter a stepwise multiple regression equation?

The list of variables that entered the equations in Bormuth's studies are listed in Table 1. Table 3 summarizes the steps in the unrestricted multiple regression program run in this project.

The stepwise regression procedure selects from a set of variables the subset that results in the best estimation of the criterion. In the present case, because the program was unrestricted, all the variables entered, but it is obvious that after Step 3 no other variable would be included at the .05

Table 2.—Zero-order Correlation Coefficients between Predicted and Obtained Mean Cloze Scores—Bormuth's Readability Formulas

Item	Obtained Cloze	Predicted Formula 1	Predicted Formula 2	Predicted Formula 3	Predicted Formula 4
Obtained Cloze	1.00				
Formula 1	.359	1.00			
Formula 2	.461	.753			
Formula 3	.495	.777	.859	1.00	
Formula 4	.374	.852	.623	.680	1.00

p (r = .373) = .01

Table 3.—Unrestricted Stepwise Multiple Correlation—Summary of Steps

Step	Variable	Standard Error of Estimate	Coefficient of Multiple Correlation	Shrunken Value	Coefficient of Determination	Significance Level
1	structure words/nouns	1.337	.466	.466	.217	.001
2	letters/sentences	1.241	.582	.569	.339	.008
3	repetitious anaphora/words	1.176	.648	.626	.419	.022
4	structure words/adjectives	1.153	.675	.645	.456	.111
5	Dale-Chall List 769/words	1.130	.701	.663	.491	.109
6	adjectives/structure words	1.084	.737	.696	.543	.044
7	personal pronouns/words	1.067	.754	.708	.569	.144
8	class inclusions/words	1.072	.760	.705	.577	.412
9	letters/syllables	1.081	.763	.699	.582	.533
10	personal pronouns/conjunctions	1.091	.765	.692	.586	.575
11	words/sentences	1.102	.768	.685	.590	.544
12	letters/words	1.099	.777	.687	.604	.296
13	syllables/words	1.112	.780	.679	.608	.606
14	letters/independent clauses	1.127	.781	.668	.610	.663
15	syllables/independent clauses	1.132	.787	.665	.620	.396
16	words/independent clauses	1.143	.791	.657	.626	.517
17	syllables/sentences	1.163	.791	.642	.626	.828
18	Dale-Chall List 3000/words	1.185	.792	.626	.627	.889
19	common noun/structure words	1.208	.792	.607	.627	.911

level of significance. This is what happened in the restricted program. After Step 7 the addition of new variables not only does not contribute substantially to the correlation, but the shrunken value of the correlation (5) starts to diminish after this point. That means that the apparent increase in the correlation is simply a spurious result that disappears when the loss of degrees of freedom is taken into consideration. If the first seven steps are compared with Bormuth's 1966 and 1969 equations, it becomes apparent that the best set of predictors differs in each case. Out of the four variables included in the 1966 equations, none entered in this equation; of the eight variables included in the 1969 equation, only three entered the present equation. On the other hand, of the eight promising variables listed in Table 1, four entered the equation. Table 4 summarizes the results and shows also the step in which the variable entered.

From the results presented in Table 4, it seems that when the full range of ability of the subjects and difficulty of materials is used, the best predictors are not the same as when a restricted range is used. This may be another reason for the reduction in the correlation values.

How well is cloze mean predicted?

Table 3 reports in column 4 the multiple correlation coefficient obtained in this study. If only those variables that reach a .05 significance level are included, the correlation is .648; that is, 42% of the total variance can be explained in terms of three variables: structural words per noun, letters

Table 4.—Variables Entering the Multiple Regression Equation Reported in 1966 and 1969 by Bormuth and Their Position in the Present Equation

Year of Study	Variable in Equation	Position in Present Eq.	Promising Variable	Position in Present Eq.
1966	Letter/independent clause	14	Syllables/independent clause	15
	Nouns/strings	19	Words/independent clause	16
	Letter/words	12	Structure words/nouns	1
	Pronouns/conjunctions	10	Adjectives/structure words	6
1969	Syllables/sentences	17		
	Syllables/words	13	Anaphora/words	3
	Personal pronouns/words	7	Class inclusions analysis/ words	8
	Dale-Chall List 769/words	5	Structure words/adjectives	4
	Dale-Chall List 3000/words	18		
	Words/sentences	11		
	Letters/words	12		
	Letters/meaningful punctuation units	2		

per sentence, referential repetition anaphora per word. Notice, nevertheless, that this is an *ad hoc* correlation, that is, an optimum value for this particular set of passages. In Study II, a cross-validation of the formula obtained in Study I is performed.

Study II

Two new sets of data, leisure-time activities and newspapers, from the same literacy study were used in this study. The passages were also 60 to 70 words long, and 10-word cloze tests had been developed on them. The subjects were again seventh graders. The same linguistic variables were computed for each passage.

For each passage five predicted cloze means were obtained: one for each of Bormuth's formulas and one for the formula obtained in Study I in this report. Zero order correlation coefficients were computed between all predicted means and the obtained cloze means.

Table 5 presents the results. All correlation coefficients are significant at the .01 level, except, again, for the results of Formula 1. Bormuth's Formula 3 gives consistently higher estimates than the other three. Formulas 2 and 3 use almost the same numerical values and variables, except that Formula 3 uses the Dale-Chall long list of common words, whereas Formula 2 uses the short list. Apparently, by including a larger range of common words, prediction can be improved substantially, at least at this grade level.

The formula developed in the textbook materials gives better cross-validation results than Bormuth's formulas. This can be accounted for in three different ways:

1. Although the materials are different in the cross-validation sample, the subjects come from the same population. Bormuth's formula was developed not only for a different set of materials, but also using subjects from a different population.
2. The best predictor variables under the present circumstances are different from when the full range is considered.
3. The third possibility is a combination of both. Although Bormuth's formulas have a greater generality, the evidence from this study tends to suggest that formulas developed for a specific population may give better results than formulas developed for a more general purpose.

Table 6 shows the intercorrelation between the estimated cloze means for the different formulas. An important result is the fact that the intercorrelations among Bormuth's formulas are higher than the correlation of any of them with the textbook formula. This again seems to suggest that under the circumstances of this study, another set of variables should be used for maximum prediction.

Summary and Conclusions

Readability formulas that have been developed for general purposes are usually cross-validated in samples of materials of a wide difficulty range and with subjects of wide ability range. In the more restricted classroom situation or when a school system evaluates materials for a specific grade level, the situation is somehow different due to the restricted range of both materials and ability level. This study was performed to test (under the latter situation) the performance of four readability formulas reported by Bormuth in 1969. These formulas were selected because

Table 5.—Zero-order Correlation Coefficients between Obtained Cloze Mean and Predicted Cloze Mean—Predictions Using Bormuth's and This Project's Formulas

	Formula				
Materials	1	2	3	4	Textbooks
Textbooks (N = 45)	.359	.461	.495	.374	.647*
Newspapers (N = 36)	.454	.428	.470	.419	.549
Recreational (N = 23)	.487	.461	.561	.525	.552
Textbook + Newspaper + Recreational (N = 104)	.410	.448	.475	.455	.495†
Newspapers + Recreational (N = 59)	.502	.495	.526	.506	.581

* Not cross-validation data

† Spurious value; includes data used for development of the formula

Table 6.—Intercorrelations between Means Predicted with Bormuth's Formulas and the Textbook Materials' Formula

	Bormuth's Formulas				
Item	1	2	3	4	Textbooks
1	1.00				
2	.926	1.00			
3	.925	.959	1.00		
4	.835	.819	.810	1.00	
Textbooks	.534	.525	.468	.472	1.00

they show an unprecedented degree of accuracy in prediction.

Sets of materials from a literacy evaluation in Madison Public Schools were used as a cross-validation sample for this study. The subjects were seventh graders, and the domains of materials were textbooks, leisure-time reading, and newspapers. The results show a reduction in the validity of the formulas; that result was expected, but a word of caution is necessary when considering the results. Because the materials and tests were not developed *ad hoc* for this study, the short length of the passages used may lead to a lower reliability of the criterion scores as well as of the linguistic variables used and/or to a curvilinear relationship among them. Since it is not possible to determine if the reduction is due to the changed circumstances of the study or to low reliability and the effect of curvilinearity, it is recommended that the results of this study be considered as provisional until another study is performed using longer passages.

The results show that a drop in the correlation should be expected from around .90 in Bormuth's cross-validation studies (2) to around .45, which can be considered as a "floor" value; that is, the validity should not drop much lower in subsequent samples. On the other hand, it seems that formulas developed in samples of subjects from the same population may perform better than general formulas. The present study also suggests that the best set of predictors may be different under the changed circumstances in which it was performed. Given the theoretical and practical foundations that Bormuth's studies have laid for a systematic exploration of readability and the more economically feasible computer technology available today, it should be possible for large school systems to develop formulas for their population that reach a high degree of accuracy even under circumstances of restricted ability and difficulty range.

Finally, the possibility could be considered of developing formulas that are specific not only to a restricted population

of subjects, but also to a universe of materials; that is specific formulas for seventh-grade textbook materials, for instance, or for newspapers and magazines. These further restrictions would improve the prediction ability of the formulas and would result in a better match between ability of the students and the difficulty of the materials.

FOOTNOTES

1. This Project was carried out in cooperation with the Curriculum Department of the Madison, Wisconsin Public School System and the Instructional Research Laboratory of the University of Wisconsin.

2. The authors wish to express their appreciation to Robert Clasen, Associate Director of the University of Wisconsin Instructional Research Laboratory for his support and counsel during this project.

3. These computation formulas are currently being revised. They may be obtained by writing Professor John Bormuth at the University of Chicago.

REFERENCES

1. Bormuth, J. R., "Readability: A New Approach," *Reading Research Quarterly*, I, 3: 79-132, 1966.
2. Bormuth, J. R., *Development of Readability Analyses,* University of Chicago, 1969.
3. Hansen, L. H.; and Hesse, K., *Results of the Assessment of Reading Literacy–An Interim Report,* Madison Public Schools, Madison, Wisconsin, 1972.
4. Klare, G. R., *The Measurement of Readability,* Iowa State University Press, 1963.
5. McNemar, Q., *Psychological Statistics* (4th ed.), Wiley, New York, 1969.
6. University of Wisconsin, *STATJOB,* University of Wisconsin (mimeograph), 1969.

TEACHER SELF-ACCEPTANCE, ACCEPTANCE OF OTHERS, AND PUPIL CONTROL IDEOLOGY

ORR N. BRENNEMAN
Cumberland Valley School District

DONALD J. WILLOWER
PATRICK D. LYNCH
Pennsylvania State University

ABSTRACT

The relationships between teacher self-acceptance, acceptance of others, and pupil control ideology were examined. Levels of self-acceptance and acceptance of others were measured using Berger's instrument. The Pupil Control Ideology (PCI) Form, based on a custodial-humanistic continuum, served as the operational definition for teacher orientations toward student control. A sample of 276 teachers responded to these instruments. Pearson product moment correlations indicated that self-acceptance was not related to PCI, but that high acceptance of others was associated with humanism in PCI. Regression analysis indicated that acceptance of others, followed by teaching level and teaching experience, predicted teacher PCI. Speculations on why self-acceptance was not associated with teacher views on control were presented.

THE CONCEPT OF SELF has long intrigued students of human behavior. William James, Freud, George Herbert Mead, and Charles Horton Cooley among others have contributed to existing thought on the subject. "Self" has been investigated from a variety of perspectives including self-actualization, self-concept, and self-acceptance. For a review see Wylie (16).

Studies by Fey (2), Sheerer (13), Berger (1), and Omwakee (11) link self-acceptance and acceptance of others. Both self-acceptance and acceptance of others often are assumed to be desirable characteristics of teachers.

The present inquiry examined the relationships of public school teachers' levels of self-acceptance and acceptance of others and their pupil control ideology. Pupil control ideology has been conceptualized on a humanistic-custodial continuum (14), and there have been a large number of investigations of both the pupil control ideology and pupil control behavior (7) of educators (15).

Studies that dealt with the relationship of various teacher predispositions and pupil control ideology report that low dogmatism (15), commitment to emergent rather than traditional values (6), low status obeisance or deference (5), high creativity (4), high levels of self-

actualization (9), and high expressed own and wanted behaviors of inclusion, control, and affection as measured by Schutz's FIRO-B scale (8) are associated with humanistic teacher pupil control ideology. In addition, high teacher sense of power was found to be related to teacher pupil control ideology–pupil control behavior congruence (12).

McAndrews (10) tested two hypotheses concerning teacher self-esteem and pupil control ideology. One hypothesis suggested that high self-esteem would be associated with a humanistic pupil control ideology. A second hypothesis proposed that teacher self-esteem would be negatively related to conformity, defined as the congruence of self pupil control ideology and the perceived pupil control ideology of colleagues operationalized in terms of "the typical teacher in your building." Neither of these hypotheses was supported empirically.

The present study builds on McAndrews' work, and refines it in the sense that it tapped a somewhat different dimension of self and added another concept. McAndrews' definition of self-esteem was based on the discrepancy between reported self and ideal self. Wylie (16) pointed out that self-esteem or congruence between self and ideal self means being proud of one's self, while self-acceptance means respecting one's self including one's recognized faults. The former concept was employed by McAndrews; this research utilized the latter concept and, in addition, the concept of acceptance of others.

Two hypotheses guided the investigation: (1) Teacher self-acceptance will be positively related to humanism in pupil control ideology; and (2) Teacher acceptance of others will be positively related to humanism in pupil control ideology.

The rationale for these hypotheses was simple and straightforward. It was grounded in the notion that an individual who is self-accepting also is likely to accept others and exhibit a humane person-oriented stance toward those with whom he or she interacts. In the case of public school teachers this type of person seems likely to hold a humanistic pupil control ideology.

Method

Instruments

In order to test the hypotheses, operational definitions for self-acceptance, acceptance of others, and pupil control ideology were required. The first two variables were measured by the Self-Acceptance and Acceptance of Others (SAAO) Form developed by Berger (1). The Pupil Control Ideology (PCI) Form devised by Willower, Eidell, and Hoy (14) served as a measure for the third variable.

For purposes of developing the SAAO Form, acceptance of self was defined as the possession of behavior patterns guided by internalized values, life-coping capabilities, sense of self-worth, and an absence of shyness or self-consciousness. Acceptance of others was defined as behavior patterns guided by acceptance of individual differences, lack of dominance, service, interest in satisfactory relationships, and a belief that individuals are responsible for their actions (1).

The instrument consists of 64 Likert-type items with five response categories ranging from "true of myself" to "not at all true of myself." The self-acceptance part of the scale is composed of 36 items with a possible range of scores of 36 to 180; the acceptance of others section contains 28 items with a 28–140 scoring range. On both scales, the higher the score, the more accepting the respondent. Reported SAAO Form matched-half reliabilities were .89 or greater, except in one case where a corrected coefficient of .75 was indicated. Validity was based on essays written by 40 subjects, with half of the subjects writing on their attitudes about self and half on their attitudes about others. These documents were evaluated by four judges and the mean ratings were correlated with scale scores. The correlations between ratings and scores were .90 for self-acceptance and .73 for acceptance of others (1).

The PCI Form taps educators' views on pupil control on a *humanistic-custodial* continuum. A humanistic orientation toward pupil control stresses an accepting, trustful view of pupils and optimism concerning their ability to be self-disciplining. A custodial pupil control ideology emphasizes the maintenance of order, distrust of pupils, and a moralistic stance toward deviance. The 20-item Likert-type device uses a 5-point response scale ranging from strongly agree to strongly disagree. Examples of items are, "being friendly with pupils often leads them to become too familiar," and "pupils can be trusted to work together without supervision" (reverse scored). The scoring range on the instrument is from 20 to 100; the higher the score, the more custodial the ideology. Reported split-half reliabilities are above .90. Validity studies were based on the use of the PCI Form, with teachers judged by their principals to be custodial or humanistic (14).

Sample

The two instruments described and an information sheet requesting demographic-type data were sent with a cover letter to the faculties of ten school building units in a single school district in central Pennsylvania. A total of 342 teachers received the packets, and 276, or 81 per cent, of them returned usable forms.

Results

The statistical method used to examine the major hypotheses was the Pearson product-moment correlation. The first hypothesis, which proposed a positive relationship between ideology, was rejected. Relevant data are presented in Table 1.

Table 1.—Correlation between Teachers' Self-Acceptance and Pupil Control Ideology

Variables	*N*	Mean	SD	r^2	r^2	*P*
Self-acceptance	276	148.2	15.6			
Pupil Control Ideology	276	52.5	9.0	−.07	.0049	NS

Separate correlations were also calculated between these two variables for the subsamples of 102 elementary teachers, 81 middle school teachers, and 93 high school teachers. None of these correlations was significant.

The second hypothesis indicated that teacher acceptance of others would be positively related to humanism in pupil control ideology. Although the correlation between the variables was only moderate, the association was a significant one, and the hypothesis could not be rejected. See Table 2 for pertinent information.

Table 2.—Correlation Between Teachers' Acceptance of Others and Pupil Control Ideology

Variables	*N*	Mean	SD	*r*	r^2	*P*
Acceptance of Others	276	111.5	10.0			
Pupil Control Ideology	276	52.5	9.0	−.28	.078	<.001

Correlations were again computed separately for teachers at the elementary, middle school, and high school levels. The correlation coefficient for elementary teachers was - .28 significant at the .01 level; for middle school teachers it was −.15 not significant; and for high school teachers it was −.35, significant beyond the .001 level.

In addition, a multiple linear stepwise regression technique was used to determine the most significant predictors of the dependent variable, pupil control ideology. The independent variables were self-acceptance, acceptance of others, teacher sex, teaching level, and teaching experience. The first variable added to the regression equation was the one which made the greatest improvement in "goodness of fit." The next most significant variable was then added until all had been considered. Those variables not maintaining a default tolerance of .01 were dropped from the equation, while those not attaining that level did not enter the equation. The final regression equation contained the variables that, in combination, represented the best predictive value while holding the other variables constant. Guilford (3) provides information on this procedure.

The results of this analysis are shown in Table 3. Acceptance of others, teaching level, and teaching experience were the most significant predictors of pupil control ideology, with acceptance of others being the single best predictor.

Table 3.—Partial and Multiple Correlations of Predictors of Pupil Control Ideology

Variables	*N*	Partial *r*	r^2	*R*	R^2
Acceptance of Others	276	−.27	.073		
Teaching Level	276	−.24	.058		
Teaching Experience	276	.19	036		
All Variables	276			.40	.16

Some additional findings of note are the following. Although male and female teachers did not differ significantly in self-acceptance, female teachers were more accepting of others than male teachers. The respective means of 113.3 and 108.7 yielded a Behrens-Fisher *t* value of 3.66 which was significant at the .001 level. Female teachers were also more humanistic in pupil control ideology than male teachers; their respective mean PCI Form scores were 51.2 and 54.7, and the resulting *t* ratio of 3.11 had a probability beyond the .005 level.

Elementary teachers, with a mean PCI Form score of 49.6, were significantly more humanistic than middle school teachers, who scored 53.9, and significantly more humanistic than high school teachers, who scored 54.5. The respective *t* values of 3.38 and 3.85 both carried probabilities beyond the .001 level. Teaching experience also was associated with a more custodial pupil control ideology. Teachers having more than five years of experience exhibited a PCI Form mean score of 54.1, while those with five years or less experience had a mean score 50.7. For this comparison, a *t* of 3.21 was significant at the .001 level. These results support those of past investigations of pupil control ideology (15).

Finally, it was found that the correlation for the entire sample between teacher self-acceptance and acceptance of others was .40. This is significant at the .001 level and is consistent with the results of previous studies.

Discussion

Our data indicate that acceptance of others—but not self-acceptance—predicts teacher pupil control ideology. Since teaching is an occupation that occurs in a setting which highlights interpersonal relationships, it is not surprising that attitudes toward others should be associated with orientations toward pupil control. However, it was not expected that self-acceptance would be unrelated to

these orientations even though our result fits McAndrews' findings (10) on teacher self-esteem and pupil control ideology.

Several speculations are suggested. The concept of self-acceptance appears to gain at least some of its theoretical utility from the focus on pathology so often found in work in social and clinical psychology. It may be that self-acceptance predicts well for those at the extremes of a self-acceptance continuum. It is also possible that those at the lower extreme of this continuum are eliminated at some point in the process of teacher selection or soon drop out of teaching in favor of some other kind of work. This is consistent with the fact that the teachers in the present sample had a higher mean level of self-acceptance than any of the groups reported on by Berger (1).

Another conjecture is that self-acceptance simply may not be as significant as has been believed in influencing job-related attitudes, especially when the socialization process functions effectively. The fact that, in addition to acceptance of others, teaching level and teaching experience were predictors of pupil control ideology, tends to buttress this contention. In contrast to self-acceptance, acceptance of others is associated with teacher pupil control orientations. This suggests that acceptance of others, as compared with self-acceptance, is less constrained by organizational and other social factors, and is the kind of personal quality that can find legitimate expression in the school's social setting.

The authors make the usual disclaimers in connection with this research. In particular, it should be borne in mind that the teacher sample came from a single school district. Nevertheless, it is believed that a number of intriguing questions have been explored both in this inquiry and in this speculative analysis.

REFERENCES

1. Berger, E. M., "The Relationship between Expressed Acceptance of Self and Expressed Acceptance of Others," *Journal of Abnormal and Social Psychology,* 47: 778-782, 1952.
2. Fey, W.F., "Acceptance of Others and Its Relation to Acceptance of Self and Others: A Revaluation," *Journal of Abnormal and Social Psychology,* 50: 274-276, 1955.
3. Guilford, J. P., *Fundamental Statistics in Psychology and Education,* 4th ed., McGraw-Hill, New York 1965.
4. Halpin, G.; and Goldenberg, R., "Relationships between Measures of Creativity and Pupil Control Ideology," paper presented presented at the Annual Meeting of the American Educational Research Association, New Orleans (mimeograph), 1973.
5. Helsel, A. R., "Status Obeisance and Pupil Control Ideology," *Journal of Educational Administration,* 9: 38-47, 1971.
6. Helsel, A. R., "Value Orientation and Pupil Control Ideology of Public School Educators," *Educational Administration Quarterly,* 7: (Winter 1971), 24-33, 1971.
7. Helsel, A. R.; and Willower, D. J., "Toward Definition and Measurement of Pupil Control Behaviour," *Journal of Educational Administration,* 12: 114-123, 1974.
8. Helwig, C., "Authenticity and Individual Teacher Interpersonal Needs," *Journal of Educational Administration,* 11: 139-143, 1973.
9. Jury, L. E., "Teacher Self-Actualization and Pupil Control Ideology," unpublished doctoral dissertation, The Pennsylvania State University, 1973.
10. McAndrews, J. B., "Teachers' Self-Esteem, Pupil Control Ideology and Attitudinal Conformity to a Perceived Teacher Peer Group Norm," unpublished doctoral dissertation, The Pennsylvania State University, 1971.
11. Omwakee, K. T., "The Relationship between Acceptance of Self and Acceptance of Others Shown by Three Personality Inventories," *Journal of Consulting Psychology,* 18: 443-446, 1954.
12. Rose, K. R., "Teachers' Sense of Power and Pupil Control Ideology and Behavior Congruence," unpublished doctoral dissertation, The Pennsylvania State University, 1974.
13. Sheerer, E. T., "An Analysis of the Relationship between Acceptance of and Respect for the Self and Acceptance of and Respect for Others in Ten Counseling Cases," *Journal of Consulting Psychology,* 13: 164-175, 1949.
14. Willower, D. J.; Eidel, T. L.; and Hoy, W. K., *The School and Pupil Control Ideology,* 2d ed., Penn State Studies No. 24, University Park, 1973.
15. Willower, D. J., "Some Comments on Inquiries on Schools and Pupil Control," *Teacher's College Record,* in press.
16. Wylie, R. C., "The Present Status of Self Theory," in Borgatta, E. F. and Lambert, W. W. (eds.), *Handbook of Personality Theory and Research,* Rand McNally, Chicago, 1969.

LEARNING DISABILITY MEASUREMENT WITH THE SYNCHROCEPHALOGRAPH

T. CHARLES HELVEY
Institute of Information and Control Systems
Tullahoma, Tennessee

ABSTRACT

This article describes a new testing method which can be used to screen learning-deficient children fast, reliably, and inexpensively out of any population of public school systems. To eliminate the difficulty of arriving at reliable individual test scores, when they are matched against the means of large statistical sample space standards, the Bayes theorem and conditional probabilities are discussed. The *Synchrocephalograph*—or Neural Efficiency Analyzer—is composed of a brain wave amplifier with highly efficient noise filters, a special purpose mini-computer, a monitoring device to eliminate artifacts, and a headset with easily applicable electrodes. It provides essentially two encephalographic parameters, which are called neural efficiency and hemispheric symmetry. The paper also describes the use, scoring, and assessment of the tests, and reference is made to the age correction in the tests.

LEARNING IMPAIRMENT can be caused either by physiological or by environmental factors. It is of great importance to know if the learning deficiency in a person is based on a functional impairment of the brain in those areas which are operative in information acquisition and recall, or if it is a result of parental, sociological, pupil-teacher relationship or other personality interactions, or drug effect.

If such distinction can be made easily, fast, and inexpensively, then a breakthrough will have been achieved by placing every child in an optimal level of the educational system for maximum learning experience. And, to comply with the law requiring the psychological screening of all school children, the proposed system will solve the difficulties by verifying the normalcy of the majority of the students, who, therefore, will not require any further diagnostic testing. By eliminating 80 to 85% of the testees, the funds and professional personnel may then be adequate to cope with the remaining workload. Thus, the professional personnel can devote intensive attention to the remedial management of those children who have learning handicaps. The information obtained with the Synchrocephalograph (SCG) is of invaluable assistance to classroom teachers, psychometrists, and parents in understanding the impairment of learning-deficient children.

The human brain capability for learning can now be measured with objectivity and reliability using the SCG. This technique does not permit a direct numerical correlation with the generally used IQ test. The reason for this is well understood and is related to the fact that the requirements for the measurements of the intelligence quotient are not in direct function with learning capability.

Use of the SCG advantageously allows for an objective measurement of the cognitive capability and related psychophysical factors of the brain; therefore, the results of this measurement provide a guidance for efficient teaching and training and most useful information for individual teaching, behavior management, and perceptual motor development. Thus, one of the major advantages of the use of the SCG is that it prevents mislabeling and misplacing a child in the educational delivery system, and proves that *learning dysfunction* can be *teaching dysfunction.*

Other applications are, for example, assistance in various manpower training programs, such as CETA, and also as a tool to approximate the objective, functional human age, thereby removing the discrepancies and injustices caused by chronological age limits in various occupations.

Theoretical Considerations

Neural Efficiency and Intelligence

Neither the mechanism nor the cybernetic interactions, which generate these systems output of the brain, is clearly understood. It is certain that these electrical activities, namely the neural information transfer rate, or neural efficiency (NE), and the hemispheric synchronization, or time-delay (TD), are independently related to general intelligence.

The electrical activity of the brain, detectable on the surface of the head, represents the statistical behavior of the neurons as they process information. This activity in the brain is analog and digital, but due to volume conduction and other technical factors, only analog signals can be measured on the surface of the head. The interaction of

many signals of different intensities and phase relations results in a particular spectrum of the composite signal. Individual differences are small in the alpha-band (8-12 Hz); however, at higher frequencies they are substantial. A large amplitude at a given frequency means that many cell assemblies, to use Hebb's terminology, are synchronized and active at that instance. Such synchronization can occur randomly, or as a necessary concomitant of the information processing program.

The concept of neural efficiency, which presently is restricted to time-domain analysis only, is based on the following testable hypotheses, as developed by Ertl:

General Hypothesis: The efficiency of information processing in the brain is related to the electrical signals required and used by the system.

Specific Hypothesis: The average frequency of non-alpha activity is related to information processing efficiency.

A great deal of effort has been spent on the IQ concept. After sixty years of work, thousands of articles, and millions of IQ tests, it is now generally agreed that we do not know exactly what human intelligence is, or how to measure it accurately. IQ test scores are not culture-free or even culture-fair, and the test scores do not successfully measure the potential to learn, but only what has already been learned. IQ test scores are also poor predictors of job success or academic achievement. There is some correlation between NE and IQ test scores at the extremes, but even less in the middle range. The results simply indicate that both methods overlap and measure some aspect of intelligence.

The brain is anatomically divided into two halves, and communication exists between the hemispheres. It is therefore reasonable to assume that the synchronization of this communication process may be an important variable in relation to intelligence. The SCG is designed to measure the degree of cross-correlation between the EEG derived from the right and left hemispheres.

Subjects with learning disabilities generally have relatively large time-difference scores, but sometimes normal NE scores. The NE variable and the symmetry variable may tentatively be regarded as the output efficiency of information management of the brain.

IQ Testing

Learning ability and efficiency are in some proportion functional to intelligence. Therefore, within the test battery which should scale the perceptual and motor development of an individual, the widely used intelligence quotient plays an enormously important role in scaling young persons.

The application of the usual, so-called "paper and pencil"-type Intelligence Quotient Test not only has the well-known low fidelity due to inherent systematic errors in the testor-testee systems dynamics, but its validity suffers further decrement if applied to individuals belonging to groups with different cultural and socio-economic background. In general, the scaling of children with respect to their learning ability is, in spite of the scientific efforts of standardization, in a most objectionable state.

If one complicates matters by asking for comparative values in learning potential of various social or ethnic subsets, such as black and white high school students of various sex and ages, one arrives at very controversial and even inflamatory issues which, so far, have not received objective, unbiased, and reliable research effort. Thus, urgent research is needed to: (1) Improve the scaling of individual children when statistically determined standards are applied; and (2) Test and apply hardware which can measure in an objective, repeatable, unbiased, fast, and inexpensive manner the learning potential of children and adolescents. Hereby various predetermined parameters such as age, sex, ethnic origin, various psychological test results, etc., would be given statistically significant sample space. Such research results could be subjected to a correlation and cross-correlation study, which would give the categorization of the population a more solid basis.

Further Limitations of Standardized Testing

The establishment of the degree of capability of a child to learn and mentally mature in function of its chronological age is a widely practiced occupation among school psychologists. Almost daily, new tests are added to the already voluminous battery. Considerable effort and funds are spent to determine the sensitivity, repeatability, and error margin of these tests when applied to a very heterogeneous population.

The inherent variability of our child population makes any standardization most difficult to achieve even in the objective tests, not to mention the more subjective tests, needed to evaluate emotional parameters. The perturbance caused by the variable influence of the testor on the testee's score is often overlooked, and in many cases the administration of a test battery to place a child precisely in a given maturity group or to determine learning deficiency is not a very reliable activity at the present time. Nevertheless, it is most important to place children into compatible groups to prevent performance decrement when exposed to heterogeneous, highly competitive groups and to uniform instruction. This, in turn, necessitates analytical evaluation through testing.

Improving Evaluative Ability

If one is confronted with a large number of test scores, the validity of each individual test in the sample space is less significant than the overall predictability of the statistically assessed values. Although the individual tests have, in most cases, a well-known error range, which might be quite significant, the aid of large numbers in the sample space will smooth out extreme values, thus rendering the test's predictability statistically acceptable. If, however, the outcome of the test score bears great importance on the future of an individual child, then the broad limits of the

scattered scores cannot be overlooked. To improve the evaluation of individual scores, which are compared to statistical averages, one can apply the Bayes theorem which is based on the principles of conditional and total probabilities.

The definition of conditional probability, namely, the probability of X occurring if Y has been already decided, can be expressed by:

$$p(X/Y) = \frac{p(XY)}{p(Y)} \qquad (1)$$

whereby $p(X) \neq 0$, and X and Y are score values in this case.

Venn-Diagrams are a good way to illustrate XY, and with such methods an adaptive algorithm can be developed by which one can give higher fidelity to the solutions in dynamic programing. This is needed for such models as the prediction of the validity of test scores; however, this method corresponds with considerable difficulties in the numerical and mathematical manipulation. Utilizing the Bayes theorem, let X_1, X_2, through X_n be the number of independent—or, rather, not interacting—variables or events whereby the probability of all X_i events must be greater than zero, and X_i should cover the total event space. The total probability of an incidental event, where $Y = X_1 Y + X_2 Y \ldots X_n Y$, is:

$$p(Y) = \sum_{i=1}^{n} \cdot p(X_i Y) = \sum_{i=1}^{n} p(Y/X_i) \cdot p(X_i) \qquad (2)$$

Assuming this statement, and utilizing Equation (1), one can obtain the probability of the event density X_j ($j = 1, 2, \ldots, n$) through the Bayes theorem, provided that the event Y has already occurred:

$$p(X_j Y) = p(X_j Y)/p(Y) = \frac{p(Y/X_j) \cdot p(X)}{\sum_{i=1}^{n} p(Y/X_i)\, p(X_i)} \qquad (3)$$

If we apply this to the prediction of the validity of test scores, then we must have some statistical prerequisites. For instance, let us assume that from a statistically significant sample space in a given test, 20% of the population score high (X_1), 20% score marginal (X_2), and 6% fail the test (X_3). The sensitivity of test is such that 90% of the failures (Y) are recognized as belonging into this class and 95% of the high scores are recognized as such. In the case of marginal scores, 50% are considered "passing" and 50% as "not passing." The problem is to find the probability of a given high scorer to be really a high scorer. With the Bayes theorem, we find that:

$$p(X_1/Y) = \frac{p(Y/X_1) \cdot p(X_1)}{p(Y/X_1) \cdot p(X) + p(Y/X_1)p(X_2) + p(Y/X_3) \cdot p(X_3)}$$

$$= \frac{0.95 \cdot 0.20}{0.95 \cdot 0.20 + 0.50\ 0.20 + 0.10 \cdot 0.60}$$

$$= 0.19/0.35$$

$$= 0.543$$

This means that only 54.3% of the high scores are really the high scores. And, for $p(X_3/Y)$:

$$p(X_3/Y) = \frac{0.10 \cdot 0.60}{0.95 \cdot 0.20 + 0.50 \cdot 0.20 + 0.10 \cdot 0.60}$$

$$= 0.06/.035$$

$$= 0.171$$

which indicates that from those who get high scores, *17.1% are failures*, and consequently, for $p(X_2/Y)$:

$$p(X_2/Y) = \frac{0.50 \cdot 0.60}{0.95 \cdot 0.20 + 0.50 \cdot 0.20 + 0.10 \cdot 0.60}$$

$$= 0.1/0.35$$

$$= 0.286$$

which means that *28.6% of the high scores are only marginal.*

As far as the failures are concerned, the conditional probabilities are:

$$p(X_1/Y) = \frac{p(Y/X_1) \cdot p(X_1)}{p(Y/X_1) \cdot p(X_1) + p(Y/X_2) \cdot p(X_2) + p(Y/X_3) \cdot p(X$$

$$= \frac{0.50 \cdot 0.20}{0.05 \cdot 0.20 + 0.50 \cdot 0.20 + 0.90 \cdot 0.60}$$

$$= 0.01/0.65$$

$$= 0.015$$

$$p(X_2/Y) = 0.50\ (0.20/0.65)$$

$$= 0.54$$

$$p(X_3/Y) = 0.90\ (0.60/0.65)$$

$$= 0.831$$

According to this, from those getting a failing score, *only 83.1% are really failing, 15.4% are marginal, and 1.5% are high scorers.*

As indicated, then, the validity of individual test scores should be based on the considerations described above. It can be proven mathematically that, by definition, one cannot expect absolute values by test scores, and their fidelity lays on an assymptote. All that one can do is to optimize the point of practical termination of the curve, which is determined by the utility of the test as well as by environmental factors and conditions which are brought about by the interactions of the subsets of the triadic system "testee-testor-test."

Naturally, by the analytical considerations of test scores, the obtained information lends itself well to treatment of

measurement if one can reduce the scores to the elementary units of information. This would enhance the comparability and, thus, the fidelity of score values, even more. If we denote the score as source of information (x), then the mean value $H(x)$ is:

$$H(x) = \sum_{i=1}^{k} P_1 \ (-\log_2 P_i) \qquad (4)$$

$H(x)$ is expressed in bits per score, of the mean logarithmic probability for all scores from a given test.

Of course, one can also make good use of human judgment by assessing hard-to-quantize test values, provided one takes the laws of comparative judgment into careful consideration. This can be stated for discriminal differences as:

$$S_1 - S_2 = X_{1,2}\sqrt{\sigma_1^2 + \sigma_2^2 - 2r\sigma_1\sigma_2} \qquad (5)$$

Hereby, S_1 and S_2 are the score values of two compared tests; whereby,

$\sigma_{1,2}$ = the sigma value, representing the proportion of judgment $P_{1,2}$. If this value is greater than 0.5, the numerical value of $X_{1,2}$ is positive; otherwise it is negative.

σ_1 = discriminal dispersion of the information of score S_1.

σ_2 = discriminal dispersion of the information of score S_2.

r = correlation between the discriminal deviation of S_1 and S_2 in the same judgment.

This type of approach is valid for experimental-analytical work where Weber's law or Fechner's law is involved, and in most other educational scaling.

Testing by Electroencephalography

Beside these considerations for the value assessment of individual test scores, hardware is now available to measure and provide numerical readout concerning the neural efficiency—that is, the learning capability—of individuals. A great deal of research has been done to relate various parameters of the electrical activity of the human brain to psychological variables. A comprehensive description of both the success and failures in this area are described by Dr. Charles Shagass in his book, *Evoked Potentials in Psychiatry (25).*

A few remarks about brain waves in general may be indicated here, as they are relevant to the Brain Wave Analyzer (BWA). Although alpha waves are most frequent in the occipital region, they do occur in the parietal and frontal lobes with frequencies between 8 and 14 Hz. Special attention should be given to the somewhat slower beta waves and their significance in brain wave analysis, as described below. Beta waves can be subdivided into beta I and beta II waves, each having different characteristics. Beta I is inhibited by increased brain activity, such as learning, while beta II waves are excited. Theta waves are also in the frequency range of the analyzer's operation, and occur during tension and frustration in the parietal region of the brain of children.

Regarding the "white noise"-type electrical activation potential, which emanates from the millions of active neurons in the brain, one would expect a less coherent and structured encephalogram. Thus, there must exist some sort of a synchronizer mechanism in the brain, the nature and location of which are unknown, which is manifest in the encephalogram. This hypothesis is strengthened by the fact that increased cerebration decreases the intensitites of most brain waves. Thus, the presence of a coordinating mechanism, governing inhibition and enhancement of the overall operation, and also its responding to certain priorities, is probable. This could account for some signatures in the EEG, although they are still too complex to permit, at present, fine-structure interpretation.

The two brain hemispheres are subdivided by the lateral fissure, which is covered by the extension of the dura. There is, however, a connection between the two hemispheres through the corpus callosum, which is probably the path of information between the hemispheres. In normal persons one side of the brain is dominant over the other. In more than 90% of the population the left side became dominant, because, for some reason, the angular gyrus region in the left half of the brain was used sooner and more frequently for learning experience. Thus, the side of the brain which gained the first start increases rapidly in potential, while the other side remains slight. However, the interpretative and many motor areas, although highly developed on one side, need the information from both sides for proper functioning. For such coherent systems output a certain, very restricted time-delay in the information transfer between the two hemispheres is mandatory. The optimal time interval is about 3 to 10 milliseconds; faster or slower time-delay has deleterious effect on cognitive functions. The BWA measures this time-delay with great accuracy.

The Synchrocephalograph

The systematic changes induced by sensory stimulation in the pattern of brain activity are known as *evoked potentials* or *evoked responses*, and it has gradually become evident that these evoked responses are very sensitive indicators of psychological and physiological states and changes in man. The study of evoked responses with the aid of modern computer technology has opened a small but significant window to the brain. Based on the work of Dr. J. Ertl and others, it appears that time-domain analysis is a useful approach toward the study of the efficiency of any information processing system, biological or electronic. Relationships have been reported (13, 17, 26) between psychometric tests of intelligence and certain time parameters of the human visual evoked responses.

There is considerable evidence that the late components of the visual evoked response are sensitive to changes in stimulus parameters involving decision making, pattern recognition, attention, and problem solving (1, 4, 6, 28). In general, there is clear and generally accepted evidence that parameters of the evoked response are related to higher levels of information processing in the brain (14).

In normal subjects the electrical activity of the two hemispheres of the brain is highly synchronized. In persons suffering from primary learning disabilities, the sychronization, however, is very poor. In most of these subjects the left-right differences are more than twice as great as in normal subjects.

Therefore, there has been developed a simple, easy-to-use, and inexpensive system to measure *neural efficiency* and *brain dysfunctions.* Considerable computer simulation and field testing have been done with this so-called Sychrocephalograph, a neural efficiency analyzer, which was designed to perform two functions:

1. To measure the rate of information transfer within the brain through the analysis of brain wave activity. The analysis yields the neural efficiency score which can be related to factors of intelligence; and
2. To measure the symmetry of the electrical activity from the two hemispheres of the brain. The degree of symmetry is related to learning disabilities and other brain dysfunctions.

Left-right difference scores are displayed by the analyzer for two major components of the evoked response. These learning difference scores will be also useful as clinical indicators in the study of learning disabilities. The results obtained with this instrument should be treated in the same careful, professional manner as the results of a medical examination or a psychiatric or psychological examination. Sometimes one obtains large asymmetry scores for many reasons besides learning disabilities. Whenever large differences are observed, medical, preferably neurological, examination is recommended to the patient. Based on the evidence available to date, the automatic Synchrocephalograph becomes a useful tool in assessing the basic neurological efficiency of the human brain, and also in the early diagnosis of brain dysfunction.

The technical description of the instrument is omitted here. However, it should be mentioned here that for maximum subject safety, the equipment is battery operated. No photic or other stimuli for evoked potentials is necessary, because the environment provides enough stimuli during the test. The testor must keep the subject relaxed because muscular tension would cause myoelectric interference with the EEG. The subject must not be given any task, such as reading, mental calculations, etc. The oscilloscope is provided to monitor an artifact-free brain wave pattern during the test, and "unclean" measurements should be discarded from the computation. The system is easy to operate, requiring only a few hours of training. Since brain waves are the basic data input it is essential that the operator be able to recognize artifacts due to improper electrode application or excessive tension on the part of the subject. This skill can be learned in a few days. The average test is completed within four minutes, which does not include the few minutes of computation and evaluation.

Validity, Reliability and the Error of Measurement

For all tests of human ability, both psychological and physiological, information about validity, reliability, and the error of measurement are important. There are some well-established rules to describe the validity of psychometric tests of intelligence. Criterion validity, i.e., a direct and independent measure of that which the test is designed to predict, is probably the most powerful demonstration of validity. From this point of view, the validity of the neural efficiency test compares very favorably with another major criterion which is based on the assumption that intelligence increases with age up to maturity. There are clear-cut developmental changes in neural efficiency. Correlation with other tests is a widely used standard of validity; however, the logic behind this method of evaluating validity is a little dubious since it assumes that another test of intelligence already has proven validity.

Impaired Levels of Consciousness

At any rate, it has been proven that in cases of impaired consciousness the neural efficiency score changes in the predicted direction, i. e., longer latencies are associated with impaired levels of consciousness. This is substantiated by the various observations concerning evoked responses and impaired levels of consciousness, examples of which follow:

1. Sleep: The latencies of the late components of the visual evoked response are increased during sleep (19, 14).
2. Experimental Delirium: Ditran, an anti-cholinergic agent, produces disruption of cognitive functioning; late components of the visual evoked response are increased in latency (3).
3. In arteriosclerotic brain syndromes, the late components of the visual evoked response are markedly prolonged (12).
4. Prolonged latencies of the visual evoked response in post-traumatic coma were found (4); the same observations were made in comatose children (13).
5. Effects of pharmacologic agents which are known to reduce levels of consciousness: A large number of preanesthetics were studied by Corssen and Domino (7, 9). They increase the latency of the components of the visual evoked response. The average latencies of all components of the visual evoked response of hypothyroid patients are longer than the latencies of a similar control group. When thyroid hormone was given to the patients, the latency differences disappeared (22). Similar studies were done with animals with the same results (16).

Age Differentiation

The relationship between age and various parameters of the evoked response has been extensively studied (1, 2, 17, 18). "Latency appears to be the evoked response characteristic that is best correlated with age. Latencies are longest during early life, shortened progressively as development proceeds, are minimal during young adulthood, and lengthened again during old age" (4). (See Figure 1.)

Correlations with Other Tests

Published correlations between the latency of various components of the visual evoked response and a number of well-established psychometric tests of intelligence range between 0.2 and 0.8 (5, 14, 26). In the majority of these studies, latencies were determined by visual inspection which introduced a considerable subjective element, with the inevitable human error, into the measurements involved. However, the Synchrocephalograph does not use any of the principles described in published reports for the measurement of the latencies of the components of the visual evoked response. No human interpretation is required, and the scores are therefore objective. Using the Synchrocephalograph, correlations of approximately 0.5 were obtained in a sample of 150 subjects. Research data are continuously flowing in from users of the instrument, and a comprehensive report will be available shortly.

Measurement Error

Due to three-digit readout, the instrumental error is less than $\mp$ 0.5%. The long-term instrument stability is about one count in 10^6 There is, of course, an inevitable measurement error, due to psychological and physiological causes, because the brain is not a machine and therefore it can not match the stability or repeatability of electronic equipment. Average variations from one measurement to the next may be of the order of 2%.

Reliability of the Analyzer

The short-term test-retest correlations range from 0.87 to 0.97 (2). Test-retest reliability coefficients one week apart are approximately 0.88. Long-term reliability of one year test-retest is 0.75.

The lower long-term reliability is probably due to maturational changes in the subjects tested. It must be noted that these reliability coefficients were obtained by cross-correlating test-retest evoked response patterns, and therefore they reflect overall stability of measures of all aspects of intelligence. Furthermore, it measures only one factor of intelligence, and this factor is also partially measured by conventional IQ tests. The neural efficiency method has the advantage that it is insensitive to language ability and is relatively uninfluenced by socio-economic and cultural factors which plague conventional IQ tests. It is probable that a multitude of environmental factors such as nutrition, early cultural enrichment, etc., have some effect on the neural efficiency score.

Evaluation and Conclusions

The evaluation of the test scores requires a simple computation based on a mathematical formula developed by Ertl (14). From the available statistically significant data

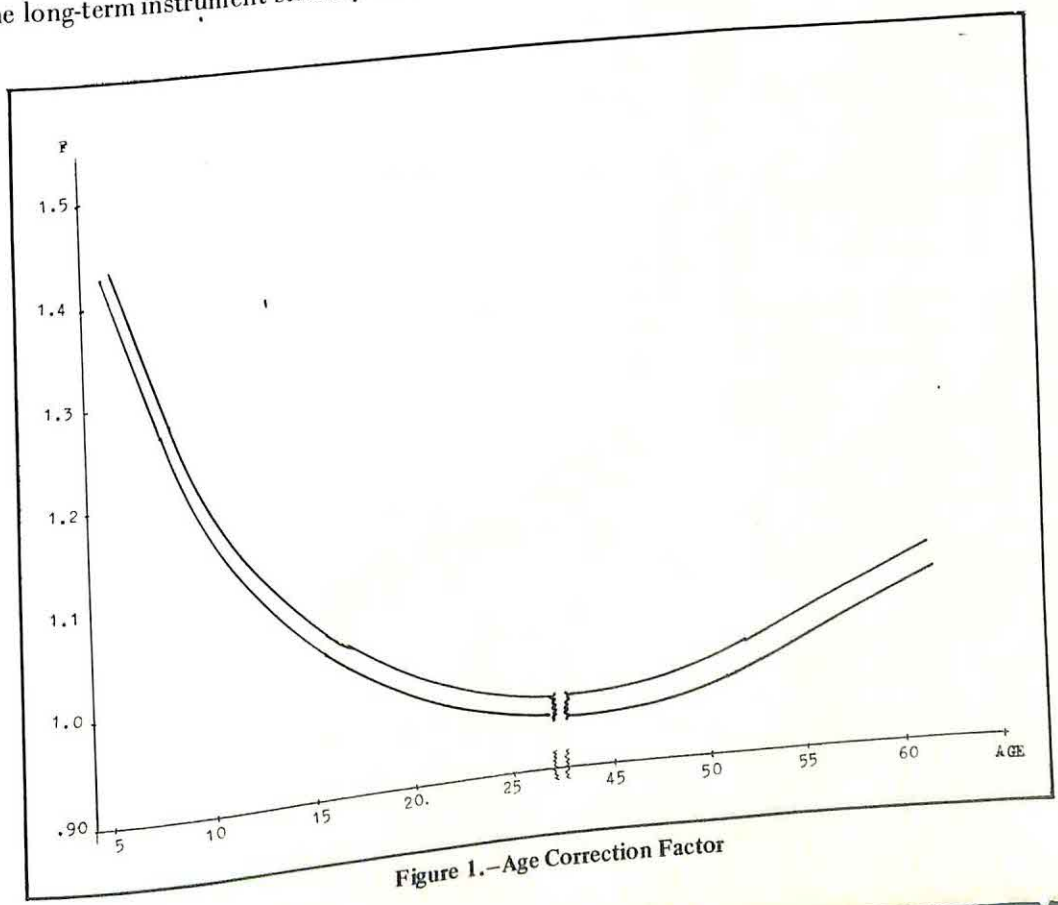


Figure 1.—Age Correction Factor

certain broad-band norms were developed. The norms for subject evaluation are shown in Figure 2.

SCORING VALUES
BWA-E-03

$$\text{Neural efficiency} = \frac{F - \text{alpha}}{1 - \text{alpha} \times \text{age}}$$

14 or less	= below average
14 - 15	= borderline
15 - 18	= average
18 - 21	= above average
21 +	= exceptional

Age correction factors:

7 or less	= .142
7 - 9	= .135
9 - 11	= .128
11 - 14	= .120
14 - 20	= .110
20 - 25	= .100
25 +	= .095

$$\text{Time difference} = \frac{\frac{1}{F} \times \text{phase}}{2}$$

0 - 10	= normal
10 - 12	= borderline
12 - 15	= below normal
15 +	= problem area

$$\text{Phase degrees} = \frac{\text{phase score}}{5.56} \text{ degrees}$$

0 - 54	= normal
54 - 65	= borderline
65 +	= abnormal

Figure 2.–Norms for Subject Evaluation

Figure 3, representing the scoring curve for neural efficiency, and Figure 4 for time-delay are quantized only with approximation, and the width of the curves indicates the inherent uncertainty. But for the task for which the equipment is presently recommended, namely, to distinguish fast and inexpensively between normal or above normal and marginal or learning-handicapped persons, a greater sophistication would not serve the purpose. The necessary age correction factor is shown in Figure 1.

The evaluation is restricted to a three-by-three matrix composed of "Brain Efficiency" with the symbolic values: Above Average, Average, Below Average; and "Time Difference" with Above Average, Average, Below Average. If either of the two parameters has a Below Average value, the subject needs special professional attention. It has been proven that as long as the brain efficiency is Average or Above Average, the time difference Below Average can easily be remediated in the classroom by removing time stress from the student. This information is one of the advantages of the SCG.

Unfortunately, the differential diagnosis provided by the SCG does not "explain" symptoms, such as dyslexia or dyscalculia in a child, or eliminate the ambiguity in the term "minimal brain dysfunction" to the degree which would be helpful in remedial prognosis.

There are also some highly interesting fringe-areas in this field needing scientific proof and further intensive research. Although there is no conclusive evidence, it is thought that the applicability of the Synchrocephalograph, possibly with multi-sensory feedback attachment, could be extended during long training to ameliorate the time delay or hemispheric asynchroniety; thus, certain learning deficiencies could be "cured." Furthermore, the SCG could be applied

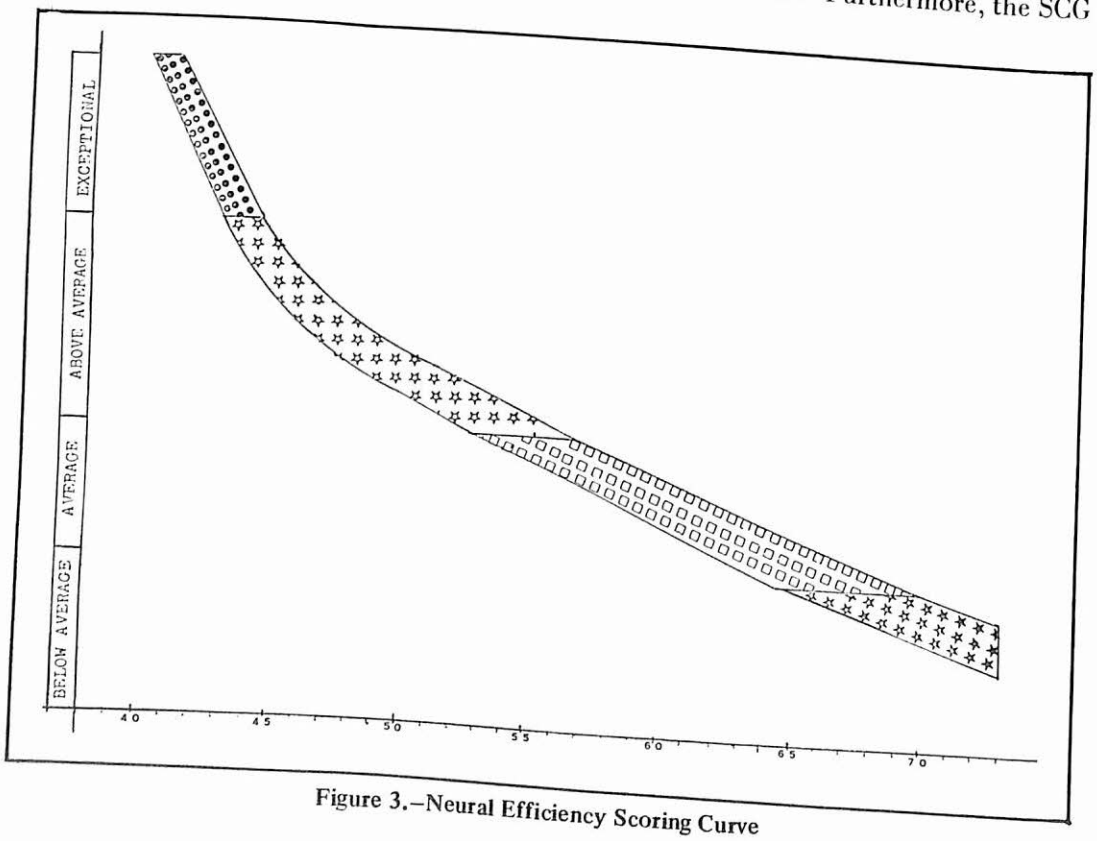


Figure 3.–Neural Efficiency Scoring Curve

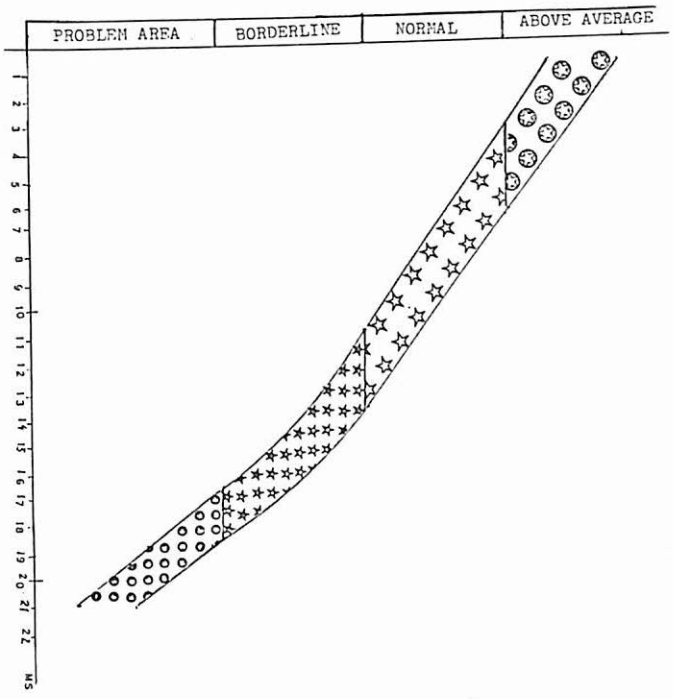


Figure 4.–Time Delay Scoring Curve

for drug abuse detection and rehabilitation monitoring, and the early detection of epileptoid and schizoid tendencies. Pending such future developments, the SCG can be successfully used in the meantime for the proven purposes as indicated in this paper.

REFERENCES

1. Beinhocker, G. D.; Brooks, P. R.; Anfenger, E.; and Copenhaver, R. M., "Electroperimetry," *IEEE Transactions on Bio-Medical Engineering*, Vol. 13, England, 1966.
2. Bradley, P. B.; Eayrs, J. T.; and Richards, N. E., "Factors Influencing Potentials in Normal and Cretinous Rats," *Electroencephalography and Clinical Neurophysiology*, 17:308, 1964.
3. Brown, J. C. N.; Shagass, C.; and Schwartz, M., "Cerebral Evoked Potential Changes Associated with the Ditran Delirium and Its Reversal in Man," in J. Wortis (ed.), *Recent Advances in Biological Psychiatry*, Vol. III, Plenum Press, New York, 1965, 223-234.
4. Callaway, E., "Averaged Evoked Responses in Psychiatry," *Journal of Nervous and Mental Diseases*, 143:80-94, 1966.
5. Chalke, F. C. R.; and Ertl, J. P., "Evoked Potentials and Intelligence," *Life Science*, 4:1319, 1965.
6. Chapman, R. M.; and Bragdon, H. R., "Evoked Responses to Numerical Visual Stimuli While Problem Solving," *Nature*, 203:155-157, 1964.
7. Corssen, G.; and Domino, E. F., "Visually Evoked Responses in Man: A Method for Measuring Cerebral Effects of Preanesthetic Medications," *Anesthesiology*, 25:330-341, 1964.
8. Creutzfeldt, O. D.; and Kuhnt, U., "The Visual Evoked Potential: Physiological Development and Clinical Aspects," in W. Cobb and C. Morcutti (eds.), *The Evoked Potentials*, Elsevier Publishing Co., Amsterdam, 1967, 29-41.
9. Domino, E. F., "Effects of Preanesthetic and Anesthetic Drugs on Visually Evoked Responses in Man," *Anesthesiology*, 28:184-191, 1967.
10. Dustman, R. E.; and Beck, E. C., "Visually Evoked Potentials: Amplitude Changes with Age," *Science*, 151:1013-1015, 1966.
11. Ertl, J. P., "Detection of Evoked Potentials by Zero Crossing Analysis," *Electroencephalography and Clinical Neurophysiology*, 18:630-631, 1965.
12. Ertl, J. P., "Information Content and the Latency of Evoked Potentials," *Final Report*, Ontario Mental Health Foundation, Grant No. 4, 1966.
13. Ertl, J. P., "Evoked Potentials, Neural Efficiency and IQ," in L. D. Proctor (ed.), *Biocybernetics of the Central Nervous System*, Little, Brown, Boston, 1969, 419-435.
14. Ertl, J. P.; and Schafer, E. W. P., "Brain Response Correlates of Psychometric Intelligence," *Nature*, 223:421, 1969.
15. Ertl, J. P., "Neural Efficiency and Human Intelligence," *Final Report*, U. S. Office of Education, Project No. 9-0105, 1969.
16. Ertl, J. P., "Evoked Potentials of Retardates," *Final Report*, Ontario Mental Health Foundation, Grant No. 180, 1969.
17. Eysenck, H. J., Letter to the Editor in *Science*, 178:7, No. 4058, 1972.
18. Fichsel, H., "Evoked Potentials," Kurze Mitteilungen aus der Medizinische, *Deutsche Medizinische Wochenschrift*, 97:309-310, 1972.
19. Helvey, T. C., "Electronic Sleep Control," *Proceedings of the First Space Congress*, 1964.
20. Kooi, D. A.; Bagchi, D. K.; and Jorden, R. N., "Observations on Photically Evoked Ocipital and Vertex Waves During Sleep in Man," *Annual of the New York Academy of Science*, 122:270-280, 1964.
21. Lille, F.; Lerique, A.; Pottier, M.; Scherrer, J.; and Thieffry, S., "Cortical Evoked Responses During Coma in Children," *Presse Medical*, 76:1411-1414, 1968.
22. Nishitani, H.; and Kooi, K. A., "Cerebral Evoked Responses in Hypothyroidism," *Electroencephalography and Clinical Neurophysiology*, 24:554-560, 1968.
23. Schuler, G.; Park, G.; and Ertl, J. P., "Low-noise Interference-resistant Amplifier Suitable for Biological Signals," *Science*, 154:1191-1192, No. 3753, 1966.
24. Shagass, C.; and Trusty, D., "Somatosensory and Visual Cerebral Evoked Response Changes During Sleep," in J. Wortis (ed.), *Recent Advances in Biological Psychiatry*, Plenum Press, New York, 1966, 321-334.
25. Shagass, C., *Evoked Potentials in Psychiatry*, Plenum Press, New York, 1972.
26. Shucard, D.; and Horn, J. L., "Evoked Cortical Potentials and Measurement of Human Abilities," *Journal of Comparative Physiology and Psychology*, 78:59-68, No. 1, 1972.
27. Straumanis, J.; Shagass, C.; and Schwartz, M., "Visually Evoked Response Changes Associated with Chronic Brain Syndrome and Aging," *Journal of Gerontology*, 20:498-506, 1965.
28. Utall, W. R., "Do Compound Evoked Potentials Reflect Psychological Codes?" *Physiological Bulletin*, 64:377-392, 1965.
29. Weinmann, H.; Creutzfeldt, O.; and Heyde, G., "The Development of the Visual Evoked Response in Children," *Archiev der Psychiatrie und der Nervenkrankheiten*, 207:323-341, 1965.

A FACTOR ANALYTIC COMPARISON OF FACULTY AND STUDENTS' PERCEPTIONS OF STUDENTS

ERNEST T. PASCARELLA
Syracuse University

ABSTRACT

In order to compare the factorial dimensions along which faculty judge students and students judge their peers, random samples of faculty and seniors from two colleges of Arts and Sciences rated the concept "Senior Students in the College of Arts and Sciences" on a 26-item semantic differential. Principal components analysis with varimax rotation yielded five faculty and six student factors. A matrix of cosines (or intercorrelations) was computed between factors to determine the degree of similarity between the two factor structures. Despite a substantial degree of overall congruence in factor structures, faculty generally associated "value" in senior students with intellectual traits while seniors associated "value" with interpersonal sensitivity.

CONSIDERABLE RESEARCH has focused on the differences and similarities between faculty and student perceptions of such variables as: teaching effectiveness (7); institutional decision-making and governance (4, 13); educational values and goals (5, 10, 14); student characteristics (3, 18); and institutional climate or functioning (1, 8, 17). The general findings of this research indicate significant group differences between the two constituencies on an extensive array of instruments purporting to measure such phenomena. Little has been done, however, to assess the degree of congruence or dissonance between the perceptual frameworks within which faculty and students view significant educational variables.

Implicit in the writing of Becker (2), Martin (14), and Tussman (19) is the suggestion that much of the cultural separation between faculty and students is due to fundamental differences in the dimensions along which the two groups structure values, goals, and perspectives. This paper reports the results of a factor analytic study of faculty perceptions of a specific student group, senior Arts and Sciences students, and those students' perceptions of their class peers. The purpose of the investigation was to determine the degree of congruence in the factors or dimensions along which both groups judge student characteristics.

Methodology

Sample

The institutions sampled in the study were two large private universities located in Central New York State with total undergraduate enrollments exceeding 10,000 students. A 20% student sample consisted of 410 senior students enrolled in the College of Arts and Sciences at both institutions. At Institution A the specific Arts and Sciences population from which the sample was drawn was 1250 seniors (52.4% female, 47.6% male). The corresponding population at Institution B was 800 seniors (33.1% female, 66.9% male). A 30% faculty sample was drawn simultaneously and randomly from Arts and Sciences faculty and the full-time equivalent of graduate teaching assistants at both institutions. The total faculty samples for Institutions A and B were 168 and 138, respectively.

Instrument

As a measure of their perceptions of senior students, seven-point semantic differential scales (15) were employed by both samples to rate the concept "Senior Students in the College of Arts and Sciences" against 26 bipolar adjective pairs. The pairs selected for use in the study were drawn

largely from two sources. The primary source was Osgood, Suci, and Tannenbaum's "thesaurus study" (15). The "thesaurus study" empirically arrays the factorial composition of an extensive number of bipolar pairs and related scales. Scales drawn from this source theoretically tapped "evaluation," "potency," "activity," and "stability" constructs. The second source was Pervin's Instrument for the Transactional Analysis of Personality and Environment (16). Scales from this source were: competitive/cooperative; pragmatic/artistic; creative/uncreative; intellectual/non-intellectual; tolerant/intolerant; and flexible/rigid. In addition, a number of scales deemed particularly relevant to the concept rated, but of unknown factorial composition, were also included (e.g., intimate/remote; unstructured/structured; supportive/frustrating; sensitive/indifferent). Such a procedure has been suggested as appropriate by Osgood, Suci, and Tannenbaum (15), provided the factor structure is empirically determined and scales of known factorial composition are also used.

Response

The instrument was distributed to the total sample at the beginning of the spring 1973 semester. The size and percentage of useable responses for both the student and faculty samples are shown in Table 1. Chi-square "goodness of fit tests" to determine the significance of differences noted between sample and population characteristics were carried out on the sex and academic major variables for students, and sex, broad area of discipline, and academic rank for faculty. The only significant chi-squares ($p < .05$) were achieved on the variable sex for faculty. At both institutions women were slightly over-represented in the faculty sample. With this singular limitation, the samples at both institutions appeared to be representative of the populations from which they were drawn.

Statistical Analysis

Analysis of the data began with an extraction of the principal components of meaning underlying faculty and students' respective ratings of the concepts on the 26 bipolar pair scales. Following Kaiser's (11) varimax criterion, components with eigenvalues ≥ 1.0 were extracted and subjected to orthogonal varimax rotation. (The rotated components will hereafter be referred to as factors.) Separate factor structures were computed for the combined faculty sample and for the combined student sample.

"Program Relate" (20) was employed to obtain a general indication of the similarity between faculty and student factor structures. "Program Relate" permits the comparison of factor structures from two independent sample groups by holding one structure fixed and rotating the

Table 1.—Sample Size and Percentage of Student and Faculty Response

		INSTITUTION A			INSTITUTION B			TOTAL BOTH INSTITUTIONS		
GROUP		SAMPLE	RESPONSE	%	SAMPLE	RESPONSE	%	SAMPLE	RESPONSE	%
SENIORS	MALE	119	46	39%	107	63	59%	226	109	48%
	FEMALE	131	75	57%	53	45	85%	184	120	65%
	TOTAL	250	121	48%	160	108	68%	410	229	56%
PROFESSORS		41	21	51%	40	14	35%	81	35	43%
ASSOCIATE PROFESSORS		36	22	61%	28	15	54%	64	37	58%
ASSISTANT PROFESSORS		33	22	67%	23	14	61%	56	36	64%
LECTURER/ INSTRUCTOR		7	6	86%	8	7	88%	15	13	87%
GRADUATE TEACHING ASSISTANTS[a]		51	21	41%	39	15	38%	90	36	40%
FACULTY TOTALS		168	92	55%	138	65	47%	306	157	51%

a - full-time equivalent

Table 2.—Matrix of Scale Intercorrelations for Faculty Responses[a]

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	*	60	66	59	62	65	54	61	63	61	64	42	42	29	12	51	74	49	17	40	60	-26	-10	60	52	49
2		**	67	67	67	65	44	69	54	59	69	43	45	23	26	66	51	54	16	51	65	-20	-08	67	57	74
3			**	69	64	70	52	77	80	69	74	54	50	20	20	63	55	61	30	61	65	-32	-22	67	65	61
4				**	67	65	61	69	65	65	67	59	56	11	21	60	53	60	25	55	64	-23	-02	64	63	64
5					**	84	57	66	57	67	72	49	49	08	12	55	50	51	16	48	63	-18	-14	84	48	52
6						**	61	66	65	64	71	56	53	10	07	56	49	51	15	55	67	-23	-21	82	52	51
7							**	49	56	51	49	53	49	02	01	39	42	36	10	37	44	-27	-08	63	35	47
8								**	74	64	74	51	46	15	20	63	56	60	23	60	58	-28	-19	68	65	57
9									**	67	69	51	45	18	11	53	52	62	20	55	61	-34	-19	61	65	49
10										**	69	51	41	16	26	58	61	59	19	53	64	-23	-14	58	61	52
11											**	54	51	19	13	59	54	61	20	56	68	-25	-07	67	57	60
12												**	78	-22	05	54	29	50	18	59	44	-25	-20	51	49	34
13													**	-23	06	51	29	47	15	49	47	-18	-23	51	41	42
14														**	15	15	25	09	03	01	22	02	11	05	18	24
15															**	21	21	21	11	14	24	-30	-02	06	31	21
16																**	43	57	26	61	55	-16	-16	50	62	62
17																	**	48	23	36	51	-16	-02	52	49	45
18																		**	35	61	52	-09	-28	49	80	45
19																			**	51	17	-02	-11	16	29	17
20																				**	61	-25	-31	47	61	47
21																					**	-33	-14	56	58	55
22																						**	36	-23	-24	-21
23																							**	-12	-27	-01
24																								**	43	51
25																									**	50
26																										**

[a]DECIMAL POINTS OMITTED

Table 3.—Matrix of Scale Intercorrelations for Student Responses[a]

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	*	74	16	39	40	31	11	23	19	31	36	05	35	-05	01	52	48	28	09	18	24	-16	06	23	-08	29
2		**	13	37	42	43	22	31	24	37	40	01	31	08	16	62	56	39	17	22	30	-19	03	23	-05	33
3			**	25	49	42	24	42	28	29	47	-06	33	27	20	51	25	43	26	34	30	-08	-29	61	35	52
4				**	42	41	34	32	18	44	48	13	37	-10	10	29	28	29	20	27	30	-24	18	25	-16	30
5					**	65	30	44	28	49	62	12	54	09	18	44	45	44	28	38	48	-13	-09	47	22	49
6						**	45	49	28	51	51	17	48	08	14	41	40	43	30	38	39	-13	-14	43	09	49
7							**	34	12	33	28	18	37	-09	07	25	10	22	29	29	41	-13	-11	23	-07	32
8								**	41	30	46	03	31	17	34	37	26	40	20	30	30	-02	-18	51	16	43
9									**	27	28	-19	03	23	37	19	17	32	18	22	37	-27	-05	32	10	25
10										**	52	11	40	07	-01	33	36	34	15	24	19	-12	-15	33	04	37
11											**	12	51	17	20	35	41	48	16	34	41	-18	01	47	14	50
12												**	37	-28	-33	00	00	07	05	20	48	-16	-15	06	-27	04
13													**	-19	-13	33	32	27	13	29	13	-06	07	35	-08	37
14														**	41	09	07	16	03	14	41	-17	02	20	40	08
15															**	07	08	30	21	19	-04	11	-11	27	34	21
16																**	52	39	19	23	18	10	-20	29	-07	34
17																	**	35	15	22	25	-17	-29	36	09	40
18																		**	29	36	26	-21	-17	41	14	41
19																			**	63	38	-20	-12	25	08	30
20																				**	05	10	-15	36	06	34
21																					**	07	-14	28	01	37
22																						-46	01	-15	05	-18
23																						**	-30	-50	-34	
24																							**	16	42	
25																								**	24	
26																									**	

[a]DECIMAL POINTS

Table 4.—Varimax Rotated Factor Loadings Derived from Faculty Semantic Differential Ratings of Seniors ($N = 157$)

VARIABLE	FACTORS I	II	III	IV	V	h^2
strong/weak (6)	.851	-.132	-.173	-.048	-.136	.792
intellectual/non-intellectual (24)	.850	-.061	-.188	-.062	.005	.770
deep/shallow (5)	.844	-.099	-.155	.004	-.035	.748
stimulating/dull (11)	.807	-.254	-.067	.130	-.019	.738
good/bad (1)	.804	-.140	.152	.038	-.119	.705
candid/deceitful (3)	.758	-.366	-.032	.154	-.197	.771
progressive/regressive (2)	.748	-.206	-.038	.347	.088	.732
active/passive (8)	.747	-.331	-.047	.159	-.147	.717
potent/impotent (4)	.746	-.275	-.211	.256	.067	.747
excitable/calm (9)	.740	-.310	.001	.045	-.264	.715
complex/simple (10)	.728	-.277	-.015	.204	-.103	.660
systematic/disorganized (7)	.719	.054	-.273	-.048	-.102	.610
creative/uncreative (21)	.701	-.234	-.017	.278	-.147	.642
intimate/remote (17)	.680	-.213	.249	.104	-.019	.581
supportive/frustrating (26)	.645	-.193	-.041	.396	.157	.635
sensitive/indifferent (16)	.570	-.446	-.172	.306	.055	.651
accepting/critical (19)	.035	-.762	.017	-.042	.024	.585
tolerant/intolerant (20)	.436	-.668	-.245	.128	-.182	.746
flexible/rigid (18)	.498	-.661	-.083	.128	-.100	.717
unstructured/structured (25)	.508	-.588	-.002	.313	-.187	.736
impulsive/restrained (14)	.268	-.065	.762	.142	.087	.684
reliable/unreliable (13)	.495	-.224	-.667	.091	-.049	.750
stable/unstable (12)	.503	-.286	-.655	.070	-.109	.780
idealistic/practical (15)	.031	-.083	.090	.861	-.110	.769
competitive/cooperative (23)	-.011	.282	.130	.103	.798	.744
pragmatic/artistic (22)	-.226	-.153	.090	-.399	.734	.780
EIGENVALUES	12.80	1.83	1.59	1.01	1.26	
PERCENT TOTAL VARIANCE	39.67	11.71	7.26	6.61	5.91	
CUMULATIVE TOTAL VARIANCE	39.67	51.38	58.64	65.25	71.16	
PERCENT COMMON VARIANCE	55.74	16.46	10.20	9.29	8.31	
CUMULATIVE COMMON VARIANCE	55.74	72.20	82.40	91.69	100.00	

NOTE: The number in parentheses after each scale indicates the number of that scale in the matrix of intercorrelations.

second structure on it until maximal similarity is achieved among the individual test vectors (test vectors in the present study are the 26 bipolar pair scales). The degree of rotation required to achieve maximal similarity is expressed as a matrix of cosines which may be regarded as a matrix of correlations between the two sets of factor variables or factor vectors derived from the two analyses (12). For the purposes of the study, two factors were considered "moderately congruent" if one factor variable or vector accounted for between 50 and 75 percent of the variance in another, as estimated by squaring the value of the cosine (i.e., cosine between .707 and .866). Factors with greater than 75 percent estimated shared variance (i.e., cosine $>$ .866) were considered "highly congruent." The reader is cautioned that since no definitive standard is available in the literature, the categories "moderately" and "highly congruent" are nominal and therefore somewhat arbitrary. Moreover, the procedure of squaring the cosine was employed to obtain a general estimate of the shared variance between factor variables. Since independent samples are used, no assumption is made that such a procedure necessarily provides the exact equivalent of the coefficient of determination obtained from single-sample data. For these reasons the use of "Program Relate" in the present study should probably be regarded more as a descriptive rather than a statistical method.

Results

Tables 2 and 3 show the respective intercorrelation matrices on which components analysis was based. Principal components analysis of faculty members' semantic differential ratings of senior students yielded five factors with eigenvalues 1.0. The composition of these five factors (i.e., the loadings of the bipolar pair scales on them) is shown in Table 4. Arrayed in Table 5 is the factor structure obtained from principal components analysis of senior students' semantic differential ratings of their class peers. As the table indicates, six factors meeting Kaiser's (11) varimax criterion were yielded by this analysis.

Table 6 displays the matrix of cosines derived from comparison of faculty and senior student factor structures. As shown in the table, a "moderate" or "high" degree of structural similarity was indicated between each of the faculty dimensions and five of the six individual student dimensions. Only Factor II in the student structure appeared to have little relationship with any of the factors in the faculty structure, thus suggesting where the most meaningful contrast between faculty and student dimensions might be made. On this factor two of the three scales

Table 5.—Varimax Rotated Factor Loadings Derived from Senior Students' Semantic Differential Ratings of Their Class Peers (N = 229)

VARIABLE	FACTORS I	II	III	IV	V	VI	h^2
stimulating/dull (11)	.753	.257	.116	.072	.102	-.100	.672
deep/shallow (5)	.706	.348	.060	.170	-.141	-.001	.673
complex/simple (10)	.683	.249	.028	.031	.113	-.044	.544
reliable/unreliable (13)	.661	.224	-.402	.126	-.007	-.109	.678
strong/weak (6)	.629	.328	.006	.314	-.091	-.088	.618
candid/deceitful (3)	.605	.047	.206	.165	-.431	.004	.624
intellectual/non-intellectual (24)	.544	.138	.180	.198	-.419	-.099	.572
potent/impotent (4)	.530	.226	.024	.193	.345	-.295	.576
active/passive (8)	.454	.152	.291	.264	-.161	-.372	.548
supportive/frustrating (26)	.446	.278	.033	.286	-.413	-.225	.579
flexible/rigid (18)	.398	.296	.259	.299	-.114	-.222	.464
progressive/regressive (2)	.198	.847	.114	.118	.109	-.112	.807
good/bad (1)	.203	.808	.017	.022	.116	-.065	.712
sensitive/indifferent (16)	.254	.731	.053	.160	.001	-.063	.632
intimate/remote (17)	.248	.728	-.010	.015	-.233	-.089	.653
idealistic/practical (15)	.015	.009	.730	.239	-.177	-.242	.680
impulsive/restrained (14)	.179	-.001	.682	-.101	-.186	.322	.645
stable/unstable (12)	.263	-.084	-.627	.215	.121	-.060	.534
excitable/calm (9)	.254	.122	.595	.193	.073	-.167	.504
accepting/critical (19)	.053	.112	.096	.849	-.086	.093	.762
tolerant/intolerant (20)	.292	.122	.075	.756	-.061	.119	.696
systematic/disorganized (7)	.395	.007	-.138	.495	.052	.166	.450
competitive/cooperative (23)	.043	-.022	.009	-.169	.847	.097	.757
unstructured/structured (25)	.145	-.073	.349	-.102	-.721	.132	.696
pragmatic/artistic (22)	-.089	-.145	-.002	.156	.028	.831	.743
creative/uncreative (21)	.538	.096	.043	.024	.069	-.604	.671
EIGENVALUES	7.91	2.84	1.83	1.50	1.36	1.04	
PERCENT TOTAL VARIANCE	18.49	12.26	8.94	8.80	8.31	6.62	
CUMULATIVE TOTAL VARIANCE	18.49	30.75	39.69	48.49	56.80	63.42	
PERCENT COMMON VARIANCE	29.15	19.33	14.10	13.88	13.10	10.44	
CUMULATIVE COMMON VARIANCE	29.15	48.48	62.58	76.46	89.56	100.00	

NOTE: The number in parentheses after each scale indicates the number of that scale in the matrix of intercorrelations.

Table 6.—Matrix of Cosines Showing the Relationship between Faculty and Student Factor Structures[a]

SENIOR FACTORS	FACULTY FACTORS I	II	III	IV	V
I	.854 (.73)	.024 (.00)	-.177 (.03)	-.139 (.02)	-.064 (.00)
II	.508 (.26)	-.081 (.01)	.166 (.03)	.154 (.02)	.241 (.06)
III	.062 (.00)	-.266 (.07)	.745 (.56)	.498 (.25)	.072 (.01)
IV	-.051 (.00)	-.880 (.77)	-.430 (.18)	.137 (.02)	.140 (.02)
V	-.053 (.00)	.322 (.10)	-.264 (.07)	.323 (.10)	.826 (.68)
VI	-.045 (.00)	-.215 (.05)	.362 (.13)	-.766 (.59)	.479 (.23)

[a]FIGURES IN PARENTHESES INDICATE ESTIMATED PROPORTION OF SHARED VARIANCE BETWEEN FACTORS

selected to tap the "evaluation" dimension (good/bad and progressive/regressive) loaded strongly with scales appearing to tap interpersonal "receptivity" or "sensitivity" (sensitive/indifferent and intimate/remote). In the faculty structure, however, the scales theoretically measuring the "evaluation" construct (good/bad, progressive/regressive, plus candid/deceitful) loaded highest on a dimension that appeared to be a measure of intellectual strength, scholarliness, and curiosity (e.g., strong/weak; intellectual/non-intellectual; deep/shallow; stimulating/dull; active/passive; complex/simple; systematic/disorganized; creative/uncreative).

Summary and Conclusions

Comparison of the factor structures obtained from faculty semantic differential ratings of seniors, and senior students' semantic differential ratings of their class peers suggested a substantial degree of overall congruence. Despite this general tendency toward factor structure similarity, however, an interesting, and perhaps significant, contrast between faculty and student structures was indicated in the scales loading on the same factor with the "evaluation" construct. In the faculty structure the "evaluative" dimension generally clustered with a dimension that appeared to measure students' intellectual and scholarly orientations; while in the student structure, "value" clustered with a "receptivity/sensitivity" dimension, which had a low correlation with scales measuring intellectual or scholarly traits.

In short, it might be suggested that faculty tend to respond favorably toward senior students largely in terms of their demonstrated intellectual capacities and orientations, i.e., those traits associated with their formal role as students and potential scholars. Seniors, on the other hand, would appear to value their peers more in personalistic and interpersonal than in formally academic or professional terms.

Considering the evidence that the vast majority of faculty-student contacts are limited to formalized classroom transactions (6), it is not particularly surprising that faculty members associate a favorable response toward students with the presence of traits and orientations which are most appropriate to this context. Moreover, as Jervis and Congdon (10:466) point out:

> Concern with the intellect and associated activities towers high in the need structures and self images of most university professors. It is only natural that they would tend to perceive and to structure their worlds in these terms.

Similarly, given the significance of interpersonal interaction in resolving the developmental tasks of maturation during college (9), it would seemingly follow that students value sensitivity and intimacy in their peers.

Beyond this, however, the fact that students in their senior year associate value with interpersonal rather than intellectual or cognitive traits perhaps suggests the extent to which the ethos of the student peer culture successfully resists faculty influence on the norms and values dominant in student life outside the classroom. Skills, orientations, and traits that facilitate one's successful functioning in the formal role of "student" (characteristics which faculty value and attempt to foster) may be relegated to a status of considerably less worth in the peer culture milieu. By not reinforcing these behaviors or orientations which faculty value, the peer culture tends to significantly decrease the likelihood of their assimilation by students even into their senior year of college.

REFERENCES

1. Bechard, J.E., *The College of Education at Michigan State University as an Organization: A Survey of the Perceptions of Its Students, Faculty and Administrators,* Unpublished Doctoral Dissertation, Michigan State University, 1970.
2. Becker, H.S., "Student Culture as an Element in the Process of University Change," in R.J. Ingham (ed.)., *Dynamics of Change in the Modern University,* Center for the Study of Liberal Education for Adults, Boston, 1966, 59-80.
3. DeColigny, W.G., *A Study of the Extent of Congruency between and among Student and Faculty Perceptions of and Reactions to Male Undergraduate Types,* Unpublished Doctoral Dissertation, Syracuse University, 1968.
4. Del Pizzo, V., *A Study of the Perceptions of Students and Faculty on Standing Campus Committees Concerning Student Participation in the Management of University of Missouri–Columbia Affairs,* Unpublished Doctoral Dissertation, University of Missouri, 1971.
5. Dunford, G.G., *A Comparative Study of What the Goals of New Mexico State University Should Be as Perceived by Students, Faculty and Administrators,* Unpublished Doctoral Dissertation, New Mexico State University, 1970.
6. Feldman, K. A.; and Newcomb, T. M., *The Impact of College on Students,* Jossey-Bass, San Francisco, 1969.
7. Hanke, J.E., *Teacher and Student Perceptions as Predictors of College Teacher Effectiveness,* Unpublished Doctoral Dissertation, University of Northern Colorado, 1970.
8. Hartnett, R. T.; and Centra, J. A., "Faculty Views of the Academic Environment. Situational vs. Institutional Perspectives," *Sociology of Education,* 47:159-169, 1974.
9. Heath, D., *Growing-Up in College,* Jossey-Bass, San Francisco, 1968.
10. Jervis F. M.; and Congdon, R. D., "Student and Faculty Perceptions of Educational Values," *American Psychologist,* 13: 464-466, 1958.
11. Kaiser, H.F., "Computer Program for Varimax Rotation in Factor Analysis," *Educational and Psychological Measurement,* 19: 413-420, 1959.
12. Kaiser, H.F., *Relating Factors between Studies based upon Different Individuals,* Unpublished Manuscript, University of Illinois, 1960.
13. Locklin, R. H.; and Stewart, C. T., "Student, Faculty and Administrative Perceptions of Decision-Making at Four Colleges," Paper presented at the Annual Convention of the American Educational Research Association, Minneapolis, March 1970.
14. Martin, W.B., *Conformity: Standards and Change in Higher Education,* Jossey-Bass, San Francisco, 1969.
15. Osgood, C.E.; Suci, G.J.; and Tannenbaum, P.H., *The Measurement of Meaning,* University of Illinois Press, Urbana, Illinois, 1957.
16. Pervin, L.A., "A Twenty-College Study of Student x College Interaction Using Tape (Transactional Analysis of Personality and Environment): Rationale, Reliability and Validity," *Journal of Educational Psychology,* 58: 290-302, 1967.

17. Pesqueria, R.E., *Comparison of Perceptions of the College Environment among Students, Faculty, Administration and Staff at the University of California,* Unpublished Doctoral Dissertation, U.C.L.A., 1969.
18. Rokeach, M.; Gladin, L.; and Trumbo, D.A., "Two Validation Studies with High and Low Dogmatic Groups," in M. Rokeach (ed.), *The Open and Closed Mind,* Basic Books, New York, 1960, 105-124.
19. Tussman, J., *Experiment at Berkeley,* Oxford University Press, New York, 1969.
20. Veldman, D.J., *Fortran Programming for the Behavioral Sciences,* Holt, Rinehart and Winston, New York, 1967.

FREEDOM OF CHOICE, TASK PERFORMANCE, AND TASK PERSISTENCE[1]

ROBERT V. KAIL, JR.
University of Pittsburgh

ABSTRACT

Much current educational literature argues that providing the learner more freedom in the learning situation enhances the learning process. This experiment tested two relevant hypotheses: (1) Ss who freely choose a task will perform better at that task than Ss who are forced to do it; and (2) Ss who freely choose a task will persist longer at that task than those who are forced to do it. The experiment employed a yoked-S design in which the first S chose to perform any of five tasks, while the second S was forced to perform that same task. The results supported only the second hypothesis. A suggested explanation of the effects of freedom of choice as a psychological variable was presented.

MUCH CURRENT literature in innovative education advocates providing the learner greater freedom to learn what he wants, when he wants, and how he wants. Such an atmosphere of freedom is presumed to support the psychological growth of the individual learner and to facilitate the learning process per se.

While there has been a plethora of claims advocating the beneficial effects of "freedom to learn," there has been a concomitant dearth of empirical research to support these claims. However, in research related to this issue the effect of the source of motivation on learning has been examined. Sims (3) demonstrated that individual motivation was vastly superior to group motivation in both reading rates and substituting digits for letters. Symonds and Chase (6) reported that in English usage tests intrinsic motivation to perform the task caused no extra learning beyond that of the practice effect of repetitive drill. However, these investigators admitted that the method used to generate "intrinsic motivation," showing the importance of correct English usage, was far from satisfactory. Kausler (2), in a study comparing the effects of ego-involvement versus task-orientation on the DuBois-Bunch Learning Test, found that the ego-involved group performed significantly better. Battle (1) compared inner-directed and other-directed junior high school students on a difficult mathematics problem and found that the inner-directed group persisted significantly longer.

In the present study the relationships between the amount of freedom given the learner in choosing a learning task and his subsequent performance and persistence on that task were investigated. It is reasonable to suspect that allowing an individual to select a task freely might yield results similar to those obtained with inner-directed, intrinsically motivated subjects. Presumably after an individual has freely chosen a task, he will be more committed to that task than an individual who is forced to perform that task. Thus, two hypotheses were tested. First, Ss who freely choose a task will perform better at that task than those who are forced to do it. Second, Ss who freely choose a task will persist longer at that task than those who are forced to do it.

Method

Subjects

The Ss were 56 undergraduate students enrolled in introductory psychology classes at Ohio Wesleyan University who volunteered to participate in the experiment in order to satisfy a course requirement.

Tasks

Five standard laboratory tasks were selected for use in this experiment. Each task was selected so that S could understand the general nature of the task based on a single sentence description. Further, each task was made quite dif-

ficult to eliminate any possible ceiling effects in measures of performance and persistence. As evidence of the difficulty of the tasks, only 1 of 56 Ss reached criterion before quitting the experiment. The five tasks used were: (1) *serial learning:* 50 common words were presented individually to S for a brief interval using a memory drum. On the first trial S simply observed the words as they were presented; on successive trials he tried to recall each word in the correct order before it appeared in the window of the memory drum. Criterion was errorless recall on two consecutive trials. The performance dependent variable was the number of words recalled correctly per trial; the persistence dependent variable was the total number of trials completed; (2) *maze learning:* S was blindfolded and given a stylus. E placed S's hand at the start of a complex maze, then at the goal. S was told he was to reach the goal as quickly and with as few errors as possible. When S reached the goal, E placed S's hand back in the start box. The criterion for this task was two consecutive trials through the maze without error. The number of errors per trial was the performance measure; total time spent on the task was the persistence measure; (3) *anagrams:* S was given a list consisting of four sets of three-five letters in scrambled order. The first three sets could be rearranged to form common English words; the letters in the fourth set could not be rearranged into any known English word. S was told to unscramble the words and write the correct word adjacent to the scrambled version. The measure of performance was the average amount of time to unscramble each meaningful word; the measure of persistence was the amount of time S spent attempting to solve the non-word anagram; (4) *motor learning:* S was told to press a handle at exactly ten lbs. of pressure. When S judged that he was pressing at this level, he told E, who recorded the level from a gauge outside S's line of sight. E gave S feedback after each press (trial), telling him the actual pressure obtained. Criterion in this task was five consecutive trials at exactly ten lbs. of pressure. The ratio of the number of correct presses to total presses was the measure of performance; the total number of trials was the measure of persistence; (5) *probability learning:* S learned the order in which a series of lights was illuminated. Because only two Ss chose this task, it was deleted from statistical analyses and is not described in detail here.

Procedure

S was given a brief description of the five tasks, rated the tasks on 7-point scales that ranged from very unappealing to very appealing, and then chose one of the tasks to perform. The first S of a pair then performed the task he had selected; these Ss constituted the free choice group. The second S also performed this task, regardless of the task he had chosen. If the second S chose the same task as the first, they were both designated as free choice Ss, and the next two Ss not choosing that task became the forced choice Ss. S was then read the instructions appropriate to the task. In addition, all Ss received the following instructions: "Please work on this task as long as it is interesting and worthwhile to you. If it becomes no longer interesting or worthwhile, please tell me and we will conclude the experiment. Are there any questions?" S then performed the appropriate task. When he had completed the task, he again rated the five tasks for appeal on a 7-point scale.

Results

The performance dependent variables for the five tasks and the associated means and standard deviations for the free and forced choice groups are presented in Table 1. There were no differences between free and forced choice Ss on the performance measure for three of the four tasks; on the motor learning task there was a tendency for forced choice Ss to respond correctly more frequently than free choice Ss (see Table 1). S's score was then converted to a z-score based on the distribution of scores for each task.[2] A comparison of all free and forced choice Ss on these z-scores indicated no differences between these groups, $t(51) < 1$.

Table 1.—Performance of Free and Forced Choice Ss on Four Tasks

		Free Ss		Forced Ss			N yoked	
		$\bar{X}$	s	$\bar{X}$	s	t	pairs	p
Task	Dependent Variable							
Serial learning	Words per trial	8.84	3.33	8.11	4.95	0.21	4	>.10
Maze learning	Errors per trial	18.40	11.50	21.00	14.79	0.27	5	>.10
Anagrams	Time per solvable word (sec.)	14.62	11.17	16.12	18.37	0.18	8	>.10
Motor learning	Correct presses/total presses	0.12	0.04	0.16	0.06	1.64	9	<.10

Table 2.—Persistence of Free and Forced Choise Ss on Four Tasks

Task	Dependent Variable	Free Ss $\bar{X}$	Free Ss s	Forced Ss $\bar{X}$	Forced Ss s	t	N yoked pairs	p
Serial learning	Number of trials	7.25	2.77	5.50	4.03	0.62	4	>.10
Maze learning	Total time (min.)	20.14	7.58	9.71	4.49	2.29	5	<.05
Anagrams	Time on unsolvable word (min.)	2.57	2.05	1.86.	1.36	0.77	8	>.10
Motor learning	Number of trials	69.67	55.95	50.22	23.12	0.49	9	>.10

Table 3.—Ratings of Tasks before and after Performance

	Free Ss			Forced Ss		
	$\bar{X}$	s	N	$\bar{X}$	s	N
Before	5.96	0.76	26	4.19	1.24	26
After	5.50	1.22	26	4.27	1.65	26

Comparable data for the persistence measures are presented in Table 2. Only on the maze learning task did the free choice Ss persist longer than forced choice Ss. However, an overall analysis using z-scores derived from the individual tasks found that free choice Ss persisted longer than the forced choice Ss, $t\,(51) = 1.88, p < .05$.

Ratings of the performed task, taken prior to and after performance of the task, are presented in Table 3. Free choice Ss preferred the task that was performed significantly more than forced Ss, both prior to, $t\,(51) = 6.80$, $p < .001$, and following, $t\,(51) = 3.02, p < .01$, performance of the task. However, ratings of free choice Ss declined following performance of the task, $t\,(25) = 1.77$, $p < .05$, while those of the forced choice Ss did not change significantly, $t(25) < 1$.

Discussion

The results of this experiment provide some support for the hypothesized relationship between freedom of choice of a learning task and success at that task. Free choice Ss did persist longer at the selected task, even though they did not learn significantly more or faster. This conflicting result is particularly surprising in light of the fact that the measures of performance and persistence are not independent of each other. In general, longer persistence should result in increased performance, merely as an artifact of the interdependence of the two measures.

That there were no differences in the learning condition does suggest a possible psychological mechanism for the effect of freedom of choice on performance. The freedom to choose may act as a source of drive and, consequently, activate the organism (5). One important result of such arousal is that the organism tends to emit the dominant response for the situation from the hierarchy of available responses. If the dominant response is the correct response, then increased drive facilitates performance. Conversely, if the dominant response is not the correct response, then an increased level of activation interferes with learning.

With the particular tasks used in this experiment, it is impossible to determine the nature of the response hierarchy on any post hoc basis. If it is assumed that on some tasks the dominant response is correct, while on other tasks it is incorrect, the inconsistent performance differences obtained would be explained. Thus it is suggested that freedom of choice increases the level of activity of an organism, and this results in increased task persistence. But, this increased drive level interacts with specific tasks and may facilitate or inhibit performance.

Such an explanation also yields certain implications for the utilization of freedom of choice in educational settings. Such freedom will be most advantageous in curricula that are designed to provide steady progress for the learner. In particular, the freedom to participate may be of great positive value for those forms of programmed instruction that utilize small steps and strive for no errors on the part of the learner (e.g., Skinner [4]).

Finally, these data may present a conservative estimate of the beneficial effects of freedom of choice. If a college student is required to participate in an experiment

for a course requirement, as these Ss were, then the free choice group is not really that, but is "free" only relative to the forced Ss. An alternative procedure would be to give S the option of non-participation as a part of the experiment. Likewise, it is doubtful whether the tasks were very interesting or motivating for S to perform. More interesting and relevant tasks that provide an accurate assessment of both performance and persistence would be necessary in future research to determine the specific manner and extent to which freedom of choice can facilitate the learning process.

FOOTNOTES

1. This research was conducted while the author was at Ohio Wesleyan University. The author thanks Harry P. Bahrick for his advice and encouragement throughout this research and helpful comments on an earlier draft of this manuscript.

2. The signs of the *z*-scores for the maze learning and anagram tasks were reversed so that a higher *z*-score represented higher performance.

REFERENCES

1. Battle, E. S., "Motivational Determinants of Academic Task Persistence," *Journal of Personality and Social Psychology,* 2:209-218, 1965.
2. Kausler, D. H., "A Study of the Relationship between Ego-Involvement and Learning," *Journal of Psychology,* 32:225-230, 1951.
3. Sims, V. M., "The Relative Influence of Two Types of Motivation on Improvement," *Journal of Educational Psychology,* 19:460-484, 1928.
4. Skinner, B. F., "The Science of Learning and the Art of Teaching," *Harvard Educational Review,* 24:86-97, 1954.
5. Spence, K. W., *Behavior Theory and Conditioning,* Yale Press, New Haven, 1956.
6. Symonds, P. M.; and Chase, O. H., "Practice vs. Motivation," *Journal of Educational Psychology,* 20:19-35, 1929.

STUDENT SELF-DISCLOSURE IN RESPONSE TO TEACHER VERBAL AND NONVERBAL BEHAVIOR

ANITA E. WOOLFOLK
ROBERT L. WOOLFOLK
Rutgers University

ABSTRACT

Eighty elementary school children were assigned to one of four experimental conditions such that each condition contained ten high self-esteem subjects and ten subjects of low self-esteem. Each group participated in a 25-minute vocabulary lesson in which the students were evaluated eight times by the teacher of the lesson. Teacher positive regard (defined as favorableness of the teacher's verbal and nonverbal evaluative communications to the students) was varied across the four conditions. Following the treatments a questionnaire was used to assess the students' willingness to self-disclose to the teacher. Statistical analyses of these data indicated that teacher positive regard was related to student willingness to self-disclose for male but not for female subjects. Congruence between verbal and nonverbal behaviors was not related to student willingness to self-disclose.

MUCH IS WRITTEN in teacher education materials about the role of the teacher as a facilitator of interpersonal relationships (2,10). She is expected to deal with students who are angry, upset, depressed, or unconcerned. In some schools she is expected to individualize instruction based upon the cognitive and affective characteristics of each child (2). As interest in "affective" education increases, the classroom teacher is charged, in addition, with the responsibility of enhancing the emotional development and self-esteem of her students (12).

It is not within the scope of this study to analyze these teacher roles or to determine their efficacy for student learning or development. Rather the focus is upon one small aspect of teacher-student interactions, variables affecting the willingness of the student to self-disclose to the teacher.

In her effort to facilitate interpersonal relationships, gather information about affective characteristics, or enhance self-esteem, a teacher may at times be hindered by the unwillingness of a student to disclose personal information. Yet the variables affecting such student disclosure have not been systematically studied. Most research on self-disclosure has examined the relationship between therapist and patient. Research in psychotherapy indicates that such therapist variables as empathy, positive regard, and congruence lead to high levels of patient self-disclosure (10). In situations other than therapy, Shapiro *et al.* (9) concluded that persons are most willing to disclose to others whom they perceive as most warm, congruent, and empathetic. The relationship between these attributes and self-disclosure has not been validated in educational settings, however.

This study systematically examined the relationship between two categories of teacher behavior (congruence and positive regard) and student willingness to self-disclose. These variables were investigated because they have been related to self-disclosure in other situations, and were defined operationally in keeping with the special characteristics of the classroom. One of these characteristics is that the teacher most frequently communicates to individuals in the entire class simultaneously. Teacher positive regard was defined, therefore, as the communication of positive evaluative statements to the class as a whole. Examples include "I'm very proud of the way you are working," "This is a very good class." Consistent with the work of Bugental (5), congruence in this study was defined as agreement among the verbal,

vocal, and facial components of spoken communication. For example, to be considered a congruent communication a teacher's positive verbal statements to the class would have to be accompanied by nonverbal indicators of positiveness such as pleasant, friendly, soft voice tone, smiling face, open arms, and relaxed posture.

The investigations of Davidson and Lang (6) and Kajita (7) of the effects of self-esteem upon interpersonal perception suggest that students' interpretations of teacher evaluative statements might be related to student self-perceptions. For this reason student self-esteem was investigated as an independent variable.

Finally, because sex differences are frequently found in studies of elementary age children on outcome variables such as achievement (1) and incidence of behavior problems (3), sex of student was investigated as an independent variable.

Method

Hypotheses

The independent variables were hypothesized to relate to student self-disclosure as follows:

1. Willingness to self-disclose will be greater for high self-esteem students.
2. Student willingness to self-disclose to the teacher will be directly related to teacher positive regard (positive evaluative communications). Willingness to self-disclose will be greatest when *both* verbal and nonverbal communications are positive, less when *either* verbal *or* nonverbal are positive, and least when *neither* verbal *nor* nonverbal communications are positive.
3. Student willingness to self-disclose will be greater when the verbal and nonverbal behaviors of the teacher are congruent.
4. Willingness to self-disclose will be related to student sex.

Subjects

The subjects were 80 fourth-grade students attending a predominantly middle-class public school in a large Southwestern city. A total of 140 fourth-graders attended the school. From this initial subject pool of 140 students, 21 children designated as "highly disruptive" by their homeroom teachers were removed. This was considered necessary to minimize teacher-student interactions other than those to be manipulated by the experimental design. Subjects were chosen from the remaining 119 students based upon self-esteem scores and sex as described in the "*Procedures*" section.

Procedures

Six weeks prior to the experiment all fourth-grade students in the school were given the Piers-Harris Self Concept Scale (8) in their homeroom classes. From the group of 119 students described above, the 20 females and 20 males having the highest self-esteem scores and the 20 females and 20 males having the lowest self-esteem scores were identified.

Five males and five females from the high self-esteem subgroup and five males and five females from the low self-esteem subgroup were then assigned randomly to each of four treatment conditions. On the day of the experiment, student absences in three of the treatment groups required last minute substitutions. Consequently the final groups had male/female ratios of 11/9, 10/10, 11/9, and 12/8. In this final selection the mean Piers-Harris Self Concept score of the 40 high self-esteem subjects was found to differ significantly from the mean for the low self-esteem group (t=18.1 $p < .001$).

One female, age 24, served as the teacher in all four conditions. She had received 28 hours of instruction, practice, and video-tape feedback in the presentation of positive and negative evaluation via the verbal and nonverbal channels. At no time during the study was she aware of the hypotheses being investigated.

For each experimental condition the 20 subjects were removed from their regular classrooms to participate in an experimental micro-lesson. In a nearby vacant classroom these students were taught a vocabulary lesson in which the teacher pronounced, wrote on the board, spelled, and used in two sentences eight different vocabulary words appropriate to the subjects' grade level. After the presentation of each word, the students were instructed to write as many "interesting and original sentences" as they could in two minutes using the vocabulary word. During this time the teacher walked around the room examining the subjects' work,but not communicating with them in any way.

Immediately after each two-minute work session and before the presentation of the next word, the teacher rendered a two-sentence evaluation of the students. The varying of these evaluations across conditions was the experimental manipulation, and all other teacher behavior was held constant.

In Condition I, the positive verbal, positive nonverbal condition, the teacher's positive verbal evaluations of the class (for example, "You're writing very interesting sentences. This must be a smart class.") were accompanied by positive nonverbal communications such as a pleasant voice tone, smiling face, open arms, and relaxed posture. Condition II presented a positive verbal, negative nonverbal teacher. Positive evaluative statements were accompanied by angry voice tone, frowning face, closed arms and rigid posture. In Condition III the teacher gave negative verbal evaluations (for example, "You're not writing very interesting sentences. This must not be a smart class.") accompanied by the positive nonverbal elements of pleasant voice-tone, et Condition IV contained only negative verbal statements and negative nonverbal communications. No student questions or comments occurred during any of the conditions. All conditions were videotaped and later rated by four judges

independently. The judges attained 100% agreement in assigning each tape to the correct condition.

In every condition, following the last evaluation, the teacher left the room and the experimenter administered to the students the instrument designed to assess willingness to self-disclose. This 16-item inventory was designed by one of the authors and is available from him. Each item is of the form "YES NO 1. If Miss ——— were my teacher, I would want to talk with her about . . . ." The endings varied from more superficial subjects (my favorite games) to more intimate subjects (what I don't like about myself). A subject's self-disclosure score was simply the total number of items to which he responded "yes."

Results

Hypotheses 1 and 2

An analysis of variance performed on the self-disclosure scores revealed no effect for student self-esteem. As indicated in Table 1, high and low self-esteem students expressed equal willingness to self-disclose to the teacher.

A main effect for teacher positive regard was found (see Table 1). Table 2 summarizes the findings of *t*-test comparisons of the means of the four conditions. As predicted by hypothesis 2, the students in the most negative condition (IV) were the least willing to self-disclose. When the mean of this most negative condition (IV) was compared with the means of each of the other conditions, the difference between the means was found to be significant beyond the .05 level for each comparison. However, Table 2 also indicates that the means of the four conditions did not become increasingly smaller as predicted by Hypothesis 2. The *t*-test comparisons revealed no significant differences among the means of Conditions I, II, and III.

Hypothesis 3

In order to determine whether congruence between teacher verbal and nonverbal behavior was related to student

Table 1.—Analysis of Variance Comparison of Student Self-disclosure Scores of High and Low Self-esteem Subjects by Four Conditions of Teacher Positive Regard (n=80)

Source	Mean Square	d.f.	F-ratio	p
Total	52.617	79		
Between	82.107	7		
Self-Esteem	.050	1	.0010	.9734
Teacher Positiveness	144.183	3	2.8982	.0400
Self-Esteem X Teacher Positive Regard	47.383	3	.9524	.4217
Within	49.750	72		

Table 2.—Means, Standard Deviations and *t*-Ratios for the Four Treatment Conditions on the Self-disclosure Instrument

Condition	Mean Score	SD	T-ratios		
			II	III	IV
I Verbal + Nonverbal +	13.3	5.40	.138	.981	2.25*
II Verbal + Nonverbal -	13.0	8.12		.939	1.76*
III Verbal-Nonverbal +	15.3	7.35			2.849*
IV Verbal-Nonverbal -	8.9	6.85			

*$p < .05$

willingness to self-disclose, the mean of Conditions I and IV combined (congruent conditions) was compared with the means of Conditions II and III (incongruent conditions) combined using a *t*-test. This difference between the means approached significance ($t = 3.61$, $p < .06$), but was in the direction opposite that predicted by Hypothesis 3. There appeared to be a tendency for subjects in the incongruent conditions to express greater willingness to self-disclose.

Hypothesis 4

An analysis of variance of self-disclosure scores by sex and teacher positive regard revealed a main effect for sex ($F=5.60$, $df=1$, p L .02) and a significant interaction between sex and teacher positive regard ($F=3.21$, $df=3$, p L .03). As depicted in Figure 1, male and female students indicated relatively equal willingness to self-disclose in Conditions I, II, and III, while the scores of the male students in the most negative condition (IV) were substantially depressed.

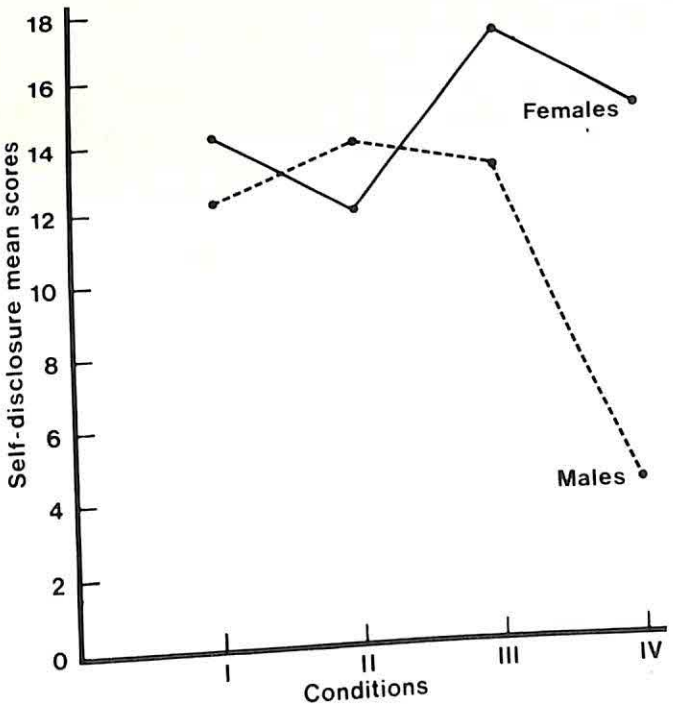


Figure 1.–Comparison of male and female students' willingness to self-disclose.

Discussion

Results indicate that congruent behavior of an unfamiliar teacher is not related to student willingness to self-disclose to her after an initial encounter. This finding was contrary to the experimental prediction and to previous research. The brevity of the micro-lesson prohibits rejecting the hypothesis that continued exposure to a teacher whose verbal and nonverbal behaviors are incongruent is related to student willingness to self-disclose. The very limited exposure of students to the teacher may have operated to establish a ceiling for self-disclosure scores, thus restricting the variability of the data. The general applicability of this finding should be regarded with some suspicion until further research provides some corroboration.

The interaction effect which was found for student sex and teacher positive regard showed that when the verbal and nonverbal evaluative behavior of the teacher was clearly negative, male students expressed an unwillingness to self-disclose. They responded "yes" to an average of only 5 out of 16 items on the self-disclosure scale. Males in the other three conditions and females in all four conditions indicated relative willingness to self-disclose, regardless of teacher evaluative behavior.

Several factors may account for these findings. It is possible that female students have learned by this age to conform more readily than males to the implicit expectations of the school. Thus for females, norms which dictate cooperation and manifestation of trust may outweight the effects of even the most negative teacher behavior upon their expressed willingness to self-disclose.

Another possibility is that this finding results from a greater willingness among students to disclose to a very negative teacher when that teacher is of the same sex. The test of this hypothesis would require another study in which male and female students were exposed to teachers of both sexes. Such an experiment would allow for the investigation of questions which cannot be addressed within the limitations of the present experimental design—those concerning the effects of sex of teacher.

Still another explanation is suggested by research on the effects of student sex on teacher punitiveness (4). It has been a clear finding that males are warned, criticized, and punished by the teacher more frequently than are females. This suggests that males may come to see unambiguous negative evaluation as a prelude to punishment. They thus would be more likely to avoid contact with teachers who might inflict punishment upon them.

Very little systematic investigation has been conducted on the role of nonverbal behavior in the classroom (14). The microteaching paradigm used in this study was chosen to provide an optimal combination of control and correspondence to classroom experience. Thus, it was not assumed that the 20-minute sample of behavior surveyed represented a typical cross section of classroom activity, but rather the closest approximation which would allow for experimental manipulation of the variables of interest. Clearly there are questions about whether one can generalize from these findings. Ultimately, any hypotheses developed by this form of research must be tested through an examination of teacher-student interactions in regular classroom settings.

REFERENCES

1. Bennett, D. A., "A Comparison of the Achievement of Fifth Grade Pupils Having Male Teachers with Those Having Female Teachers," *Dissertation Abstracts*, 27: 4032-4033, 1969.

2. Biehler, Robert, *Psychology Applied to Teaching,* 2d ed., Houghton Mifflin, Boston, 1974.
3. Bower, E. M., "A Process for Identifying Disturbed Children," in Dupont, Henry (ed.), *Educating Emotionally Disturbed Children: Readings,* Holt, Rinehart and Winston, New York, 1969.
4. Brophy, J.E.; and Good, T.L., *Teacher-Student Relationships: Causes and Consequences,* Holt, Rinehart and Winston, New York, 1974.
5. Bugental, D. E.; Kaswan, J. W.; Love, L. R.; and Fox, M. N., "Child versus Adult Perceptions of Evaluative Messages in Verbal, Vocal, and Visual Channels," *Developmental Psychology,* 2:367-375, 1970a.
6. Davidson, H.H.; and Lang, G., "Children's Perceptions of Their Teachers' Feelings toward Them Related to Self-perception, School Achievement, and Behavior," *Journal of Experimental Education,* 29:107-118, 1960.
7. Kajita, E., "Self-esteem, Affect and Interpersonal Cognition," *Japanese Psychological Research,* 10:111-122.
8. Piers, E. V., *Manual for the Piers-Harris Children's Self concept Scale,* Counselor Recordings and Tests, Nashville, 1969.
9. Shapiro, J. G.; Kraurs, H. H.; and Truax, C. B., "Therapeutic Conditions and Disclosure beyond the Therapeutic Encounter," *Journal of Consulting Psychology,* 16: 290-294, 1969.
10. Truax, C. B. and Carkhuff, R. R., *Toward Effective Counseling and Psychotherapy,* Aldine, Chicago, 1967.
11. Weigand, J. E., *Developing Teacher Competencies,* Prentice-Hall, Englewood Cliffs, N. J., 1971.
12. Weinstein, G.; and Fantini, M., *Toward Humanistic Education: A Curriculum of Affect,* Praeger Publishers, New York, 1970.
13. Woolfolk, R.L.; and Woolfolk, A.E., "Student Responses to Teacher Verbal and Nonverbal Behavior," *American Educational Research Journal,* 1974, in press.
14. Woolfolk, R.L.; and Woolfolk, A.E., "Nonverbal Teacher Behavior: A Rejoinder," *American Educational Research Journal,* 1974, in press.

BEHAVIORAL COMPONENTS OF SCHOOL READINESS

TIMOTHY M. FLYNN
Southern Illinois University

ABSTRACT

The purpose of the study was to determine the characteristics in the preschool migrant child that were related to school readiness. The relationship between change in school readiness and behavioral changes was investigated using 61 male and 72 female migrant preschool children (CA = 3.9 to 4.9). Change during a three-month interval in a compensatory preschool program in self-concept, delay of gratification, self-control, and risk taking was analyzed using multiple linear regression. For the girls, growth in self-concept and delay of gratification was significantly related to growth in school readiness, while for the boys, measures of change in self-control predicted growth in school readiness.

READINESS AS defined by Ausubel (1) refers to "the adequacy of existing capacity in relation to the demands of a given learning task" (p. 246). Maturation according to Ausubel has a different and much more restricted meaning. It represents development that takes place independently of prior learning. A child's level of school readiness depends not on maturation alone but on varying proportions of maturation and learning (1). Thus a major component of the child's lack of readiness for school is due to gaps or inadequacies in his prior learning.

Difficulty occurs when readiness is confused with the concept of maturation. In the present study a measure of the child's intellectual maturation at the beginning of an enrichment program was statistically controlled in order to separate the child's school readiness due to prior learning from his school readiness due to intellectual maturation during the program. Specifically, the present study is concerned with the personality characteristics of the migrant child which are related to school readiness due to experience.

In contrast with previous research, the present study is concerned with achievement due to behavioral characteristics rather than the child's intellectual development.

Thus, determination of the nonintellectual characteristics of migrant children that are involved in development of the child's readiness for academic material was the purpose of this study. Removing the role the child's intellectual development (cognitive skills) plays on his readiness for school eliminates any confounding relationship that may exist between personality and intelligence. Mischel (10), for example, states that most personality measures relate more highly to intellectual measures than to different measures of the same trait.

It was hypothesized that development in the migrant child's self-concept, in his ability to delay his gratification, exercise self-control, and take moderate risks would be associated with growth in school readiness. Support for this hypothesis is provided by research using not only preschool children but research using older children and adults. The characteristics of children who display more achievement behavior were studied through a longitudinal investigation (12) based on standardized tests and ratings of children's behavior in nursery and elementary school as well as in the home. The authors determined that girls whose IQ's increased during the preschool years were able to delay gratification of their desires until some future time. Mischel (9) also found a significant relationship between preference for immediate smaller, or delayed larger, reinforcement in choice situations, and "*n*" Achievement (responses to pictures scored for achievement motive). Mischel's sample consisted of 112 Trinidadian children between the ages of eleven and fourteen years. Haggard (5) found that high achievers had better self-control than equally gifted children who were not achieving at such a high level.

Support for a relationship between self-concept and achievement was provided by Crandall, Katkovsky and Preston (4), who assessed the amount of time elementary school-age children chose to spend in intellectual activities during free play time while at a summer camp. Boys who predicted their own success in intellectual activities spent more time engaged in intellectual activities, while no such relationship was obtained for the girls in the study. McClelland (8) examined the relationship between "*n*" Achievement (response to pictures scored for achievement motive) to risk taking in 26 children in kindergarten and 32 children in third grade. In both groups of subjects, individuals with high "*n*" Achievement tended to take moderate risks, while students with low "*n*" Achievement preferred either safe or speculative enterprises.

In contrast with the reviewed studies which examined behaviors independently of one another, the present study examines the characteristics acting together in predicting achievement. This makes it possible in the present study to determine whether the characteristics are independent of one another. Due to sex differences found in the reviewed studies, a separate analysis was conducted for each sex.

Method

The behavioral measures used in the study were administered three months apart by four black female psychometrists. The subjects were 132 four-year-old (three years, nine months—four years, nine months) black migrant children (71 girls and 61 boys) participating in a program of compensatory education being conducted for preschool children of migrant workers in south Florida. Total enrollment in the program was 300 students, but pre- and post-attendance of less than three months. It was thought that most in the South due to the nature of their migrant lives. These children are representative of the rural culturally deprived. Possibly, these children are even more deprived than most in the South due to the nature of their migrant lives.

The pre-test measures were administered approximately two weeks after the child entered the program, with the post-tests administered three months later. All of the instruments administered to measure the variables under study are available from the Educational Testing Service of Princeton, New Jersey. Several of the measures are experimental devices lacking in validity and reliability data. A brief description of each of the measures follows the behavior they measure.

Cognition

The cognition measure included the sum of correct responses on the following three measures. These measures are thought to tap the major components of cognitive development in preschool children.

ETS Matched Pictures Comprehension Task measures listening and recognition of word and sentence properties. The measure was developed to meet the need for a series of syntactically structured tasks which would require minimal responses from the child (i.e., pointing). The tasks consist of a "Matched Picture" presentation of 20 cards containing pairs of stimulus pictures. Both pictures contain similar elements, but they depict different relationships. The examiner asks the child to point to the similar elements.

ETS Story Sequence Task, Part II measures speaking, retelling, comprehension, and creative speech. The test materials consist of two sets (three and four cards each) of cartoon style sequences using animals as characters. The examiner tells the subject to listen carefully to the story because the subject is to repeat the same story. The subject's version of the story is recorded on tape for later scoring and interpretation.

Matching Familiar Figures measures the child's reflection-impulsivity tendencies. The subject is shown a set of four pictures, then a single standard. His task is to identify the one comparison figure among the four that is identical to the standard by pointing to the correct figure.

Delay of Gratification

The Mischel Technique measures ability to delay gratification. The subject is shown two rewards (candy) and is

told that he can have the smaller one now or the larger one at some later period (specified by the Examiner). He is asked whether he wishes the smaller or the larger of the two items.

Risk Taking

The first task in the Risk Taking measure consists of showing the child two bags. The child looks into the first bag and sees a toy (car) in it. He is told that the other bag may be empty or may have five toys in it. The child is then asked if he would rather have the car or the other bag. If the child selects the bag, the game is over. If the child selects the car, then he is shown the contents of the bag and is asked to choose another bag. The same choice situation is again presented to the child. If when first asked he selects the bag that may have five toys in it, he receives two points; however, if he selects the bag the second time, he receives one point.

Self-concept

Brown IDS Self-concept Referents Test measures the child's perception of self. The procedure involves taking a photograph of each subject to use in asking the subject questions about his picture. Each positive response receives a score of one; each negative response receives a score of zero. The child is asked to respond with a "yes" or "no" as to whether the child in the picture is "happy," "clean," "ugly," "talks a lot," "good," "scared," and so forth.

Self-control

Motor Inhibition measures the child's self-control. The child performs two motor acts: drawing a line between two points and walking a distance of six feet. He practices each act and then is timed as he performs it as slowly as he can. The child's score is the time it takes the child to complete the drawing. The Motor Inhibition Ability Test was introduced by Maccoby, Dowley, Hagen, and Degerman (7).

Dependent Variable

Cooperative Preschool Inventory (CPI) measures general knowledge, listening for word meaning and comprehension, writing (form copying), speaking, and quantitative skills. The CPI was designed as an assessment procedure for use in individual testing of children age three to six years (3). The CPI consists of 85 items which were selected on the basis of a principal components factor analysis. Williams and Stewart (13) reported a reliability of .93 (coefficient α) for a sample of 445 children attending a summer Head Start program. The author obtained a coefficient α reliability of .88 for the CPI administered to 191 migrant preschool children.

Statistical Procedure

Multiple linear regression analysis (2) was used to examine the relationship between growth in the four traits and growth in school readiness not accounted for by the intellectual development (cognition) upon entering the compensatory program. The analysis conducted examined the relationship between the growth in each trait and the growth in school readiness with all traits acting together in predicting achievement.

The pre-test measures on the characteristics being studied were used as covariates to the post-test measures instead of using gain scores. This approach provided a more reliable measure than the use of gain scores. The measure of cognitive development (combination of three ETS tests) was taken with the other pre-tests.

Results

Regression analysis allowed the investigator to estimate the proportion of criterion variance that can be accounted for by the complete system of the five traits (including cognition) and by each individual trait. An examination of the unique contributions of the five variables to the prediction of achievement was made by comparing the complete system R^2 value to the prediction system with one of the variables omitted. It is an estimate of the independent contribution of the omitted variable and may be evaluated by means of the F statistic.

The variables used in the regression analysis were: (a) cognition (variable 1); (b) delay of gratification (variables 2, pre-test and 3, post-test); (c) risk taking (variables 4, pre-test and 5, post-test); (d) self-concept (variables 6, pre-test and 7, post-test); (e) self-control (variables 8, pre-test and 9, post-test); and (f) the criterion variable, the post-test CPI score (Y) with the pre-test CPI measure used as covariants (variable 10).

On examination of Table 1, it is apparent that only self-control for the boys accounted for a significant proportion of the achievement growth variance (6.5 percent). This finding indicates that a positive relationship between growth in achievement and growth in self-control exists for preschool migrant boys attending a compensatory program. Results for the girls attending the program indicate that change in delay of gratification and self-concept accounted for a significant proportion (4.7 and 23.2 percent, respectively) of achievement growth variance.

Table 2 presents the pre- and post-test measure and standard deviations for the measures used in the study. Examination of the means revealed that the mean of the girls' delay of gratification post-test measure was lower than the mean of the pre-test measure. This indicated that the girls who were less likely to delay gratification on the post-test measure than on the pre-test measure were the ones who gained the most in achievement.

Discussion

The finding with the most statistical and potentially practical importance was the relationship between growth in self-concept for girls. Sears (11) found that boys were

Table 1.—Regression Equations for Full Prediction System and for Full Prediction System Minus Post-test Measure for Each of the Four Variables Believed To Be Associated with School Readiness

Boys (N=61)

10-Variable Prediction System (R^2=.649)

$$Y= .001x_1 + .675x_{10} - .277x_6 - .048x_2 - .016x_8 - .039x_9$$
$$.005x_7 - .009x_3 + .244x_9 + .000x_5$$

9-Variable Prediction System with Risk Taking Omitted (R^2=649)

$$Y= .007x_1 + .675x_{10} - .277x_6 - .048x_2 - .016x_8 - .039x_9$$
$$.005x_7 - .009x_3 + .244x_9$$

9-Variable Prediction System with Self-Control Omitted (R^2=.604)

$$Y= .014x_1 + .714x_{10} - .230x_6 - .075x_2 + .068x_8 - .039x_9 +$$
$$.000x_7 - .030x_3 - .007x_5$$

9-Variable Prediction System with Delay of Gratification Omitted (R^2=649)

$$Y= .006x_1 + .677x_{10} - .273x_6 - .048x_2 - .017x_8 - .040x_9 +$$
$$.006x_7 + .244x_9 + .00x_5$$

9-Variable Prediction System with Self Concept Omitted (R^2=.649)

$$Y= .007x_1 + .675x_{10} - .277x_6 - .047x_2 - .016x_8 - .039x_9 -$$
$$.009x_3 + .244x_7 + .000x_5$$

Girls (N=72)

10-Variable Prediction System (R^2=.699)

$$Y= .179x_1 + .506x_{10} + .051x_6 - .082x_2 - .050x_8 - .113x_9 +$$
$$.550x_7 - .236x_3 + .154x_9 - .050x_5$$

9-Variable Prediction System with Risk Taking Omitted (R^2=.696)

$$Y= .174x_1 + .509\ x_{10} + .038x_6 - .085x_2 - .045x_8 - .116x_9 +$$
$$.554x_7 - .232x_3 + .152x_9$$

9-Variable Prediction System with Self-Control Omitted (R^2=.681)

$$Y= .188x_1 + .514x_{10} + .010x_6 - .074x_2 + .035x_8 + .134x_9 +$$
$$.546x_7 - .230x_3 - .046x_5$$

9-Variable Prediction System with Delay of Gratification Omitted (R^2=.651)

$$Y= .141x_1 + 535x_{10} + .045x_6 - .006_2 - .303x_8 - .090x_9 +$$
$$.533x_7 - .144x_9 - .041x_5$$

9-Variable Prediction System with Self-Concept Omitted (R^2=.469)

$$Y= .777x_1 + .360x_{10} + .202x_6 + .025x_2 + .000x_8 - .333x_9 -$$
$$.208x_3 + .151x_7 - .100x_5$$

Table 2.—Boys (N=61) and Girls (N=72) Pre- and Post-test Means and Standard Deviations

Trait	Boys				Girls			
	$\bar{X}$		SD		$\bar{X}$		SD	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post
Self-Concept	9.36	13.86	2.53	2.64	9.49	13.61	2.73	2.68
Delay of Gratification	.30	.44	.24	.22	.57	.43	.30	.33
Risk Taking	1.58	1.64	.61	.68	1.82	1.81	.63	.64
Self Control	28.32	44.54	7.88	8.23	26.59	34.03	8.06	7.45
Cooperative Pre-School Inventory (CPI)	29.62	42.40	9.42	10.01	35.70	48.83	8.67	9.08
Cognitive Development	19.82		7.48		25.24		7.27	

more accurate than girls in evaluating their ability. The girls in the present study may have become more accurate because of the opportunity provided by the preschool program to compare their achievement with those of other children. Boys, because of their initially more accurate evaluations of their achievements, would not be as likely to increase in the accuracy of their self-concept during the program as would the girls.

The negative relationship between delay of gratification and gain in preschool readiness supports Maccoby's (6) theory of a curvilinear relationship between activity level and achievement. Extreme activity or inactivity, according to Maccoby, would result in few or too fleeting contacts with experience, respectively, but that towards some midpoint there would be sufficient interaction with the environment, at an appropriate "rate," for learning to take place. Thus there is an optimum point on the activity dimension, and this is identical for both boys and girls. Partial support for this position can be found in Table 2, where after three months in the compensatory program, the girls' and boys' post-test delay of gratification means are identical. The finding of a curvilinear relationship as predicted by Maccoby (6) was not possible in the present study because the Mischel measure of delay of gratification is a single binary response.

While the negative relationship between delay of gratification and school readiness for girls is interesting, a replication of these findings will be needed before any major change in the preschool program should be considered. The other findings have been supported by previous research and indicate a need to examine the nursery school curriculum to determine how best to develop these characteristics. Development of these affective characteristics of the preschool child should complement development of the child's cognitive skills. It would be unrealistic, for example, to have a unit on self-concept development. Instead, every activity the child engages in during the program could be an opportunity for the child to gain self-esteem. This will require the teacher and aides to examine their behavior with the children to determine what changes are necessary to promote these characteristics. When making lesson plans for an art project, for example, the teacher should consider along with her major objectives for the project how to use this opportunity to encourage and reward the development of impulse and control in the boys and at the same time promote more "outgoingness" in the girls.

These suggestions are tentative, yet they do deserve application and study. It is suggested that differential treatment on the basis of sex in the area of impulse control be applied to nursery school children in a controlled setting. If the preschool children in the experimental class show more gain in achievement than the control class, then substantial support will be provided for the speculations presented. Only then can we infer a causal relationship between impulse control and achievement.

REFERENCES

1. Ausubel, D. P., "Viewpoint from Related Disciplines: Human Growth and Development," *Teacher College Record,* 60:245-254, 1959.
2. Bottenberg, R. A.; and Ward, J. E., "Applied Multiple Linear Regression," 6570th Personnel Research Laboratory, Aerospace Medical Division, Air Force Systems Command, Lackland Air Force Base, Texas, 1963.
3. Caldwell, B. M., *The Preschool Inventory Technical Report,* Educational Testing Service, Princeton, N.J., 1967.
4. Crandall, V.; Katkovsky, W.; and Preston, A., "Motivational and Ability Determinants of Young Children's Intellectual Achievement Development," *Child Development,* 33:643-661, 1962.
5. Haggard, E. A., "Socialization, Personality and Achievement in Gifted Children," *School Review,* Winter Issue:318-414, 1957.

6. Maccoby, E. E., *The Development of Sex Differences,* Tavistock, London, 1967.
7. Maccoby, E. E.; Dowley, E. M.; Hagen, J. W.; and Degerman, R., "Activity Level and Intellectual Functioning in Normal Preschool Children," *Child Development,* 36:761-770, 1965.
8. McClelland, D., "Risk Taking in Children with High and Low Need for Achievement," in J. Atkinson (ed.), *Motives in Fantasy, Action. and Society,* D. Van Nostrand Co., Princeton, N.J., 1958, 306-321.
9. Mischel, W., "Father Absence and Delay of Gratification, Cross-cultural Comparison, *Journal of Abnormal and Social Psychology,* 63:116-124, 1961.
10. Mischel, W., *Personality and Assessment,* Wiley, New York, 1968.
11. Sears, P. S., "Correlates of Need Achievement and Need Affiliation and Classroom Management, Self-concept, Achievement and Creativity," unpublished manuscript, Laboratory of Human Development, Stanford University, Stanford, Calif., 1962.
12. Sontag, L. W.; Baker, C. T.; and Nelson, V., "Mental Growth and Personality Development: A Longitudinal Study," *Child Development Monograph,* No. 2, 1958.
13. Williams, R. H.; and Stewart, E. E., "Some Characteristics of Children in the Head Start Program," *Project Head Start, Final Report,* Section One, Educational Testing Service, Princeton, N.J., 1966.

ON THE SEPARATION LEVEL OF GRADES ON A MULTIPLE-CHOICE EXAMINATION

M. A. HAMDAN
R. G. KRUTCHKOFF
Virginia Polytechnic Institute and State University

ABSTRACT

The separation level $S(Z_1,Z_2)$ between two grades Z_1 and Z_2 on a multiple-choice examination, as introduced by Krutchkoff (1), was based on the normal approximation to the distribution of W, the number of answers (out of N) known by the student. Thus, the conditional probability function of W given Z was derived by Krutchkoff in a series form involving the standard normal distribution function; and the resulting tabulated values showed higher probabilities at both tails of the distribution. In the present paper, an accurate closed form for $P(W/Z)$ is derived which permits calculations to be performed on a programmable desk calculator. This closed form is based on an exact distribution of W as binomial with parameters N and p_1, the proportion of subject matter known by the student. Hence, we recalculate values of $P(W/Z)$ and $S(Z_1,Z_2)$ for Krutchkoff's example.

THE SEPARATION LEVEL of grades on a multiple-choice examination as a quantitative probabilistic criterion for correct classification of students by the examination was introduced by Krutchkoff (1). The separation level $S(Z_1,Z_2)$ of two grades Z_1 and Z_2 with $Z_1 < Z_2$ is the probability that a student with grade Z_2 knows a greater proportion of subject matter than a student with grade Z_1. Krutchkoff's derivation of $S(Z_1,Z_2)$ is based on the normal approximation to the distribution of W, the number of answers (out of N) known by the student, with the total normal probability for $W < 1$ accumulated at $W = 0$ and the total normal probability for $W > N - 1$ accumulated at $W = N$. As a result of this approximation, the conditional probability function of W given Z, $P(W/Z)$, (Table 3 of Krutchkoff [1]), showed higher probabilities at both tails of the distribution. Another approach, mentioned by Krutchkoff (1) but not employed, is to use a renormalized normal distribution truncated at $W = 0$ and $W = N$. The approach to be considered here seems more realistic than this latter approach and leads to simple closed form solutions.

In the present paper, the separation level is derived based on an exact distribution of W as a binomial random variable with parameters N (the total number of questions on the examination) and p_1 (the proportion of subject matter known by the student). It turns out that $P(W/Z)$ can be obtained in a closed binomial form which can be directly calculated. The present results can be applied to Krutchkoff's data (Table 1 of reference 1) and accurate values of $P(W/Z)$ and $S(Z_1 Z_2)$, as given in Tables 1 and 2 of the present paper, can hence be recalculated.

Derivation of $P(W/Z)$ and $S(Z_1, Z_2)$

Let us set: the number of answers known = W, the number of correct guesses = Y, the total number of correct answers = Z; so that

$$Z = W + Y. \tag{1}$$

Our derivation of the conditional probability function of W given Z is based on the following two assumptions:

1. The probability distribution of W is binomial (N,p_1), where p_1 is the proportion of subject matter known by a student.
2. The probability distribution of Y is binomial $(N - W, p)$, where $p = 1/n$, and n is the number choices for each f the N questions.

It follows that

$$P(w) = Pr(W=w) = \binom{N}{w} p_1^{w} q_1^{N-w}, \quad w=0, 1, \ldots, N; \quad q_1 = 1 - p_1 \tag{2}$$

$$P(z/w) = Pr(Z=z/W=w) = \binom{N-w}{z-w} p^{z-w} q^{N-z}, \quad z = w, w+1, \ldots, N; \; q = 1-p. \tag{3}$$

Hence, we have

$$P(z) = Pr(Z=z) = \sum_{w=0}^{z} P(z/w)P(w). \tag{4}$$

It can be directly verified that

$$\binom{N}{w}\binom{N-w}{z-w} = \binom{N}{z}\binom{z}{w} \tag{5}$$

and hence Equation (4) takes the form

$$P(z) = \binom{N}{z} p^z q^{N-z} q_1^N \sum_{w=0}^{z} \binom{z}{w} \left(\frac{p_1}{pq_1}\right)^w = \binom{N}{z} p^z q^{N-z} q_1^N \left(1 + \frac{p_1}{pq_1}\right)^z, \quad z = 0, 1, \ldots, n \tag{6}$$

By Bayes' inversion formula, Equations (2), (3), (5) and (6), we thus obtain

$$P(w/z) = P(w)\,P(z/w)/P(z) = \binom{z}{w}\left(\frac{p_1}{pq_1}\right)^w \left(1 + \frac{p_1}{pq_1}\right)^{-z} = \binom{z}{w} P^w Q^{z-w}, \quad w = 0, 1, \ldots, z \tag{7}$$

where $Q = 1 - P$, and

$$P = \frac{p_1}{p_1 + pq_1} = \frac{np_1}{1 + (n-1)p_1}, \text{ since } p = \frac{1}{n}. \tag{8}$$

Equation (7) shows that the probability function of W given Z is, in fact, binomial with parameters z and p as defined in Equation (8).

Estimation of p_1 and P

By Equation (3), we have

$$E(Y/w) = (N-w)p, \text{ and hence by Equation (2)} \tag{9}$$

$$E(Y) = (N - Np_1)p \tag{10}$$

It follows that

$$E(Z) = E(W) + E(Y) \tag{11}$$

$$= Np_1 + (N - Np_1)p, \text{ so that}$$

$$p_1 = \frac{E(Z) - Np}{N(1-p)} \tag{12}$$

Thus, estimators of p_1 and P are

$$\hat{p}_1 = \frac{n\bar{Z} - N}{N(n-1)} \text{ and } \hat{P} = \frac{n\hat{p}_1}{1 + (n-1)\hat{p}_1} \tag{13}$$

Application

Upon applying the above results to Krutchkoff's (1) example, the following results are obtained:

$$N = 60, n = 4, \bar{Z} = 25.85, \hat{p}_1 = .2411, \hat{P} = .5596.$$

Table 1.–The Conditional Probability Function $P(W/Z)$

W \ Z	F(16)	D(21)	C(26)	B(32)	A(39)
0					
1					
2	.0004				
3	.0023	.0001			
4	.0095	.0005			
5	.0290	.0022	.0001		
6	.0675	.0076	.0005		
7	.1225	.0206	.0019	.0001	
8	.1751	.0459	.0058	.0003	
9	.1978	.0842	.0148	.0010	
10	.1759	.1284	.0320	.0028	.0001
11	.1219	.1631	.0592	.0072	.0003
12	.0646	.1727	.0940	.0160	.0009
13	.0252	.1520	.1287	.0314	.0024
14	.0069	.1103	.1518	.0541	.0056
15	.0012	.0654	.1543	.0824	.0118
16	.0001	.0312	.1348	.1113	.0224
17		.0117	.1008	.1331	.0386
18		.0033	.0640	.1410	.0599
19		.0007	.0343	.1320	.0841
20		.0001	.0152	.1090	.1069
21			.0055	.0791	.1229
22			.0016	.0503	.1278
23			.0004	.0278	.1200
24			.0001	.0132	.1017
25				.0054	.0775
26				.0018	.0530
27				.0005	.0324
28				.0001	.0177
29					.0085
30					.0036
31					.0013
32					.0004
33					.0001

All omitted entries are zero to at least four significant places.

Table 2.–Separation Levels $S(Z_1, Z_2)$

Z_1 \ Z_2	D(21)	C(26)	B(32)	A(39)
F(16)	.7765	.9435	.9931	.9995
D(21)		.7501	.9413	.9937
C(26)			.7747	.9546
B(32)				.7928

Hence, we calculate by Equation (7) the binomial probability function $P(w/z)$ for z = 16, 21, 26, 32 and 39. The results are given in Table 1, to be compared with the results of Table 3 of Krutchkoff (1).

Now, the separation level of z_1 and $z_2 (z_1 < z_2)$ is

$$S(z_1, z_2) = Pr(W_1 < W_2 / z_1 < z_2)$$
$$= \sum_{i=1}^{z_2} P(W_2 = i/z_2) \sum_{j=0}^{\min(i-1, z_1)} P(W_1 = j/z_1)$$

Using Table 1, we calculate the values of $S(z_1, z_2)$ given in Table 2.

Conclusions

By using a binomial probability mass function for the known (without guessing) number of questions, a very simple formula is obtained for $P(W/Z)$. Although this does not significantly alter the previous conclusions concerning the exam, it does permit the calculations to be performed on a programmable desk calculator.

REFERENCE

1. Krutchkoff, R. G., "The Separation Level of Grades in a Multiple-Choice Examination," *The Journal of Experimental Education*, Volume 36, No. 1, Fall 1967.

RANDOM RESPONSE TECHNIQUES FOR REDUCING NON-SAMPLING ERROR IN INTERVIEW SURVEY RESEARCH

BARBIKAY BISSELL POHL
San Francisco State University

NORVAL FREDERICK POHL
University of Santa Clara

ABSTRACT

For "sensitive-area" questions in interviewing, random response techniques can be useful in establishing the types of questions to be asked and the methods by which the respondent can confidentially provide answers. This paper discusses three randomized response techniques (the dichotomized question, the unrelated question, and the single question/random answer), the simple probability theory that makes the techniques work, and some questions about the practical application of these techniques by the educator.

RESEARCHERS ARE well aware of the pitfalls of non-sampling error, particularly when personal interview techniques are used. Surveys on human populations have established that an interviewee's refusal to respond to "sensitive-area" questions or his intentional giving of incorrect answers to the questions are more likely the rule rather than the exception. Several reasons for the giving of non- and/or incorrect answers by the interviewee have been advanced: (1) modesty; (2) fear of being thought bigoted; (3) reluctance to admit to unlawful or socially deviant behavior; and (4) reluctance to confide intimacies to strangers.

For whatever the reasons, many individuals fail to answer interviewers' questions. This so-called "non-cooperative" group (12:235-272) includes two types of individuals and results in two types of non-sampling error: the non-answerer who creates "refusal bias" (2:355-361; 12:261-269), i. e., failing to respond, and the incorrect-answerer who creates "response bias" (7:280-325), i. e., purposely providing incorrect answers. These types of bias error can be quite serious. Cochran (2:235-245) has advised researchers to scrutinize their methodology with the intent of avoiding non-sampling error and has warned that the presence of such error in the data is likely to result in misleading or even incorrect conclusions.

Intuitively, the problems of refusal bias and response bias become most serious as respondents are questioned about matters they perceive as "sensitive" or whenever truthful answers may place the respondent in an unfavorable light. The experiences of field researchers show that while controversial assertions seem to elicit resistance from the interviewee, innocuous questions typically receive rather full cooperation and truthful answers (13:63).

When interviewee resistance to a question is anticipated (and, unfortunately, it often isn't simply because its existence is so "personally" defined), the usual strategy is to provide special training to sensitize the interviewer. Typically, the interviewer is taught to anticipate resistance, build a close rapport with the interviewee, create an open atmosphere and a feeling of acceptance of ideas, and hear and see (body language) what the interviewee is "really saying" (11:574-587; 4:537-543). Such training of an interviewer is not only prohibitive in cost in most cases, but also questionable in terms of the extent to which it actually reduces non-sampling error (8:161). There seems to be a natural reluctance on the part of most interviewees to confide certain feelings or facts to anyone—let alone a stranger—particularly if the responses to the questions are recorded on paper or on audio-visual tape with (or without) name and address identification.

Researchers today are also finding that questions which demand answers "too revealing" can and will be challenged both on ethical and practical grounds. Respect for a "right

to privacy" is demanded in public surveys as elsewhere (14: 884).

The traditional role of the interviewer has also changed. Today, the interviewer is keenly aware that he does not have the confidentiality privileges of the lawyer, doctor, or priest, but yet has the responsibility not to betray the trust which is rightfully expected of him (5:520-521). Also, the interviewer may contribute to non-sampling error simply by his reluctance to ask sensitive questions and thus may omit questions (contributing to refusal bias) or alter the questions asked (contributing to response bias) (9).

To reduce non-sampling error and at the same time protect the anonymity of the interviewee and the confidentiality of the interviewer, Warner (13) has devised an interviewing method called the randomized response technique which is based on answering probabilistically selected questions.

The Randomized Response Technique

In its original form, the randomized response technique has the respondent in the interview situation answer one of two questions in a designed set without revealing to the interviewer which question has been answered. The pair of questions is so structured that each of the questions could receive the same classification of answers (e. g., both questions can be answered true/false, or yes/no); and, thus, the identity of the question is not revealed by the nature of the answer. Also, the two questions are worded such that the response is not necessarily incriminating. For example, the two questions might be of the form: (1) Are you a member of Group A?, and (2) Are you a member of Group *not* A? This type of question might be designed to gain information from a group of student respondents as to their personal experience with homosexuality in a college dorm. The two questions might then be: (1) "Have you had at least one homosexual experience in a college dorm during this past year?" and (2) "Have you had no homosexual experiences in a college dorm during the past year?"

Data Collection

The respondent would be allowed to confidentially and randomly select one of the two questions to answer. For confidentiality, only the answer he gives is recorded—no indication is made as to which question was answered. Note that a "true" or "false" answer is appropriate to either question and thus the interviewee, looking at the pair of questions, believes he has retained his privacy even though he responds accurately to either of them.

The randomizing process on choice of question is a controlled one, however. A typical procedure would involve giving a spinner (or other randomizing device) to the respondent with the direction that he confidentially spin it and answer the question number to which it points. The spinner may be marked such that 70% of the time it would "land on" or indicate that Question 1 should be answered and 30% of the time it would land on or indicate that Question 2 should be answered.

Given the proportion of time that the respondents will be directed by the spinner to answer Question 1 or Question 2 (and assuming the respondents cooperate and answer truthfully), simple probability can be used to estimate the actual number or percent of individuals belonging to Group A or, in the example above, those who have had at least one homosexual experience in a college dorm during the past year. Statistically,

$$P(\text{true}) = P(\text{randomly selecting Question 1 and answering "true"}) + P(\text{randomly selecting Question 2 and answering "true"})$$

or

$$P(\text{true}) = P(Q1 \text{ selected}) \times P(\text{true}/Q1 \text{ selected}) + P(Q2 \text{ selected}) \times P(\text{true}/Q2 \text{ selected})$$

If we let P_1 = the probability that $Q1$ is selected,
π_1 = the probability that $Q1$ is answered "true,"

then $P(\text{true}) = P_1\pi_1 + (1 - P_1)(1 - \pi_1)$. Because the P_1 value of the randomization device is pre-set before the interview starts, π_1 becomes the only parameter to be estimated.

Suppose that for a randomly selected group of 400 students (sophomores, juniors, and seniors) from a predominantly "live-in" church-related school, the randomized response technique as described above resulted in a total of 176 of the 400 students (or 44%) responding "true" (to Questions 1 *and* 2). Assume that the randomizing mechanism was pre-set such that there was a 70% chance that a student would select (by having the pointer land on) Question 1 and a 30% chance that the student would select (by having the pointer land on) Question 2. Using the 44% total "true" response (or, .44) in this sample as an estimate of the proportion of the population that would answer, "true," or $P(\text{true}) = P(\hat{\text{true}}) = .44$, and the 70% chance that a student would select Question 1 as the probability of selecting that question, or $P_1 = .7$, then

$$P(\hat{\text{true}}) = P_1\hat{\pi}_1 + (1 - P_1)(1 - \hat{\pi}_1)$$

$$.44 = .7\hat{\pi}_1 + .3(1 - \hat{\pi}_1)$$

$$.44 = .7\hat{\pi}_1 + .3 - .3\hat{\pi}_1$$

$$.14 = .4\hat{\pi}_1$$

$$\pi_1 = .35$$

Therefore, the estimate of the true proportion of students having at least one homosexual experience in a college dorm during the last year ($\hat{\pi}_1$) is 35%.[1]

Under the assumption that all "true" and "false" responses are truthful, Warner (13:64-65) has shown that the expected value, or $\Sigma\hat{\pi}_1$, of the proportion of "true" answers in the sample is the maximum likelihood estimate of the true population proportion π_1. Also, because the variance

of the estimate is easily calculated[2], the construction of confidence intervals and/or hypothesis testing is quite straightforward using either the binomial distribution or, with the usual large sample sizes, the asymtotic normal approximation.[3]

Selection of P-*values*

The rationale for using a randomized response technique is based on the assumption that the procedure will elicit better cooperation (fewer refusals and/or fewer intentionally incorrect answers) from interviewees. If P_1 were set equal to 0 or 1, the whole randomizing process would degenerate into the traditional procedure of simply asking the sensitive question. At the other extreme, if P_1 were set equal to .5, the interviewee would, in fact, be furnishing no information. For *P*-values set between 0 and .5 or between .5 and 1 (exclusive), the interviewee is providing only probabilistic information as to his group membership. As the *P*-values approach .5 from either direction, less and less information is gained from the respondent as larger estimate variances result.

The question of the sample size required given a desired level of precision depends on the *P*-value selected. Presumably, the closer *P* is set to .5, the more likely respondents are to cooperate in answering the "pair" of questions. That is, the less information requested, the more likely the respondent is to cooperate. However, as the *P*-value approaches .5 from either direction, the variance of the estimate of the population parameter π(VAR π) approaches a maximum. Thus, the real issue involves selecting a *P*-value close to 0 or 1 so as to minimize the necessary sample size, yet far enough away from 0 or 1 to assure cooperation from the interviewee.

It is obvious that if all interviewees told the truth, the randomized response technique would require a larger sample size than that required for the traditional approach, for any desired degree of precision in the estimate. However, the more important comparisons are between the randomized response technique and traditional interview techniques under the realistic assumption that the traditional estimates are biased due to less than 100% truthful reporting. Warner (13) has presented evidence to this point and has shown that with even minimal untruthful reporting, the randomized response technique can out-perform the traditional interview technique.

The Unrelated Question Randomized Response Technique

The randomized response model, as outlined in the previous section of this paper, involves an interviewee answering one of two questions in a pair—the questions being related to the extent that they are direct opposites of one another. Field researchers (10) have suggested that this dichotomization of the questions in the pair may actually be confusing to many interviewees in the sense that the second question may involve a double negative and thus may be perceived to be of the form, "Heads, you win; tails, I lose." The actual result of dichotomizing the questions may be quite different, therefore, than that intended. While the researcher sees the dichotomization of questions to mean that *neither* a "true" or a "false" response should be stigmatizing, the interviewee may feel that *both* are. The supposed psychological advantage of dichotomizing the questions may, then, actually be perceived as a disadvantage.

To overcome such a limitation of Warner's randomized response technique, the unrelated question randomized response technique was developed by Simmons (5). The strategy of this technique is to have the respondent answer one of two randomly chosen questions, as he would in the previous technique, only this time the questions are unrelated instead of dichotomized. For example, the two questions might be of the form: (1) Are you a member of Group A?, and (2) Are you a member of Group B?, where membership in Groups A and B are unrelated and membership in only one of the groups is stigmatizing. A specific question might be designed to gain information from a group of student respondents as to their personal experience with drug experimentation. The two questions might then be: (1) "Do you smoke marijuana at least once a week? " and (2) "Do you own a car?"

Data Collection

The respondent would be allowed to confidentially and randomly select one of the two questions to answer, as in the previously discussed technique. Again, note that a "yes" or "no" answer is appropriate to either question and the interviewee presumably believes that he has retained his privacy even though he responds accurately to the question he chooses to answer.

Again, also, a controlled randomizing device is used. The selection mechanism will be set to indicate Question 1 a certain proportion of the time and Question 2 the complementary proportion.

The supposed advantage of this unrelated question approach is that the interviewee can see more clearly that the two questions from which he chooses are entirely different. However, a disadvantage with this approach is that only "yes" responses are implicating from the interviewee's point of view.

As initially developed, the unrelated question randomized response technique involved two unknown parameters: the proportion of the population who were actually Group A members, and the proportion of the population who were actually Group B members. In the example here of the questions on marijuana smoking and car ownership, the two unknown parameters are: (1) the proportion of the population who smoke marijuana at least once a week, and (2) the proportion of the population who actually own a car. The probability formula for this model would thus be,

$$P(\text{yes}) = P(\text{randomly selecting Question 1 and answering "yes"}) + P(\text{randomly selecting Question 2 and answering "yes"})$$

or

$$P(\text{yes}) = P(Q1 \text{ selected}) \times P(\text{yes} \mid Q1 \text{ selected}) + P(Q2 \text{ selected}) \times P(\text{yes} \mid Q2 \text{ selected})$$

If we let P_1 = the probability that $Q1$ is selected, and
π_1 = the probability that $Q1$ is answered "yes," and
θ_1 = the probability that $Q2$ is answered "yes,"
then $P(\hat{\text{yes}}) = P_1 \hat{\pi}_1 + (1\text{-}P_1)\hat{\theta}_1$. Although the value for P_1 is known, values for both $\hat{\pi}_1$ and θ_1 are unknown.

To solve for two unknowns, it is mathematically necessary to have at least two equations. Thus, it appears that the unrelated question randomized response technique would require that two samples be taken. The resulting set of simultaneous equations would be of the form[4],

$P(\hat{\text{yes}}) = P_1 \hat{\pi}_1 + (1\text{-}P_1)\hat{\theta}_1$ for sample No. 1, and
$P(\hat{\text{yes}}) = P_2 \hat{\pi}_1 + (1\text{-}P_2)\hat{\theta}_1$ for sample No. 2

The requirement of two samples is a serious drawback to using the unrelated question randomized response technique. However, it may be possible to select the unrelated question so that the probability of a "yes" response, or θ_1, is known beforehand and thus need not be estimated. For example, the question set might be: (1) "Do you regularly 'cheat' when taking classroom exams?" and (2) "Were you born in the month of January?"

Presumably, information regarding the proportion of births in the month of January would be easily attainable from census data. Thus, when the proportion of respondents answering "yes" to the unrelated or non-sensitive question can be determined *a priori* or exogenously, only one sample will be needed to estimate the one unknown parameter.

However, the parameter value secured from the census data, etc., is appropriate only so long as it is valid for the population under study (i. e., the target population). As an example, the birth months of elementary and high school students may reflect recent trends toward "planned parenthood" and/or "spaced childbirth" and thus may differ from patterns found in the United States population as a whole.

Selection of P-*values*

The purpose of Simmons' unrelated question randomized response technique was to secure better interviewee cooperation. That is, Simmons believed that randomizing responses was good but that Warner's dichotomized question approach was often confusing to the interviewee. Also, to insure truthful reporting with Warner's model, P-values close to .5 (low information content per interviewee) would presumably have to be used, whereas with clearly unrelated questions, P-values close to 0 or 1 (high information content per interviewee) can be used.

The statistical implication of Simmons' questioning strategy (which allows for P-values being close to 0 or 1) is that relatively small sample sizes can result in relatively good predictions, i. e., small variance estimates. Greenberg *et al.* (5) has shown, in fact, that under rather general conditions, Simmons' unrelated question technique will out-perform Warner's model. This is generally true when only one sample is needed (only one parameter is unknown), but may also be true under certain conditions when two samples are needed (two parameters are unknown).

The Single Question/Random Answer Randomized Response Technique

A third approach to the randomized response technique has been suggested by Greenberg *et al.* (5) Unlike either Warner's dishotomized questions or Simmons' unrelated questions, Greenberg's method involves only the single sensitive question. The randomness of response is gained by the opportunity for the interviewee to select from three possible answers. The question in this technique would be of the same general type as in the other two techniques, i. e., of the form: Are you a member of Group A?, where membership in Group A is stigmatizing. A specific question might be designed to gain information from a group of student respondents as to their personal feelings about minority student treatment on the school campus. The question might then be: "Do you feel that favoritism in terms of 'easy grades' is given to minority students?"

Data Collection

The respondent would select an answer to the question based on the concealed outcome of some randomizing device. For example, a spinner showing three area designations—red, white, and blue—might be used. The interviewee would be instructed to answer honestly the sensitive question if the spinner showed red; answer "no" if the spinner showed white; and answer "yes" if the spinner showed blue. The probability model for this technique becomes,

$$P(\text{yes}) = P(\text{randomly selecting red—the sensitive question—and answering "yes"}) + P(\text{randomly selecting blue})$$

or

$$P(\text{yes}) = P(\text{red selected}) \times P(\text{yes} \mid \text{red selected}) + P(\text{blue selected})$$

If we let P_1 = the probability that red is selected, and
P_3 = the probability that blue is selected, and
π_1 = the probability that the sensitive question—red—is answered "yes,"

then $P(\hat{\text{yes}}) = P_1 \hat{\pi}_1 + P_3$. Note that only "yes" answers to the question are incriminating, and that this would be a similar drawback for the interviewee as would his "yes" response in the unrelated question technique. However, this single question model has the advantage for the researcher of always requiring only one sample since only one parameter needs to be estimated (π).

Selection of P-values

The single question/random answer randomized response technique was originally designed for the purpose of overcoming one of the shortcomings of Simmons' unrelated question method, i. e., the necessity to estimate (or know) the population proportion associated with the neutral question. And, to this end, the single question technique is effective.

However, interviewees may find the format of the single question technique confusing. Similar to the concern associated with Warner's dichotomized question technique, interviewees may believe the question to be a trick. The fact that the method, in two out of its three possible conditions, actually tells a respondent what type of answer to give, may make it suspect. Thus, it appears that the initial issue in using the single question technique is convincing the interviewee of the method's honesty. Once this "trust relationship" has been accomplished, the question of selecting the *P*-value so as to best insure cooperation and minimize sample size requirements can be addressed.

There appear to be two generalizations concerning the selection of *P*-values for Greenberg's single question/random answer model: the *P*-values associated with "red," "white," and "blue" must all be greater than 0; and because only "yes" responses can be stigmatizing, the probability associated with "blue" must be relatively high.

Summary, Conclusion, and Implications

Two types of non-sampling error in interviewee research are "refusal bias" and "response bias." These errors often result from respondents giving no answer or a purposely incorrect answer to a "sensitive area" question.

A randomized response technique is useful to the interviewer and interviewee in assuring both persons that the answer made to the incriminating question is confidential and, thus, is more likely to be truthful. The *dichotomized question* variation by Warner allows the interviewee to respond to his "random" selection of one of a pair of directly opposite questions. The *unrelated question* variation by Simmons allows the interviewee to respond to his "random" selection of one of two unrelated questions. The *single question/random answer* variation by Greenberg *et al.* allows the interviewee only the one incriminating question, but a choice of three responses. In each of these cases, the selection of the question to answer (or the response to make) is accomplished with the help of a randomizing device (e. g., a spinner, a box of marbles, a die).

The various randomized response techniques for obtaining answers to sensitive questions are intuitively appealing. For the most part, the theoretical-statistical issues have been resolved. The following questions of application and methodology, however, persist:

There are no rules or guidelines for identifying which technique variation should be used under what circumstances.

The critical issue of selecting optimal *P*-values has not been resolved. At best, a researcher can recognize a necessary tradeoff between obtaining interviewee cooperation (insuring anonymity) and working within situational constraints of maximum sample size.

The "best" type of randomizing mechanism seems to be open to debate. To date, researchers have tried spinners, plastic boxes filled with colored balls, decks of cards, and random number tables with varying results.

The correlation between the understandability of a method and the cooperation it achieves has not been studied. In other words, to what extent does the interviewee (and interviewer) have to understand the underlying randomizing features of the model to insure his (their) cooperation? Does the degree of sophistication associated with probability theory preclude the use of randomizing techniques or certain audiences? Would it be necessary to carefully explain the purpose and mechanism of randomization to each interviewee?

Perhaps further research will lead to definitive answers to these methodological or application questions.

FOOTNOTES

1. Note that when $P_1 = .5$, the equation is unsolvable.
2. Warner (13) has shown that VAR $\hat{\pi}_1 = 1/n\ [1/16(P_1 - ½)^2 - (\hat{\pi}_1 - ½)^2]$
3. It is possible for π to take on values outside the 0 to 1 range. For example, suppose $P_1 = .7$ and the sample produces $P(\text{true}) = .25$. The solution equation becomes

$$.25 = .7\hat{\pi}_1 + .3 - .3\hat{\pi}_1$$
$$-.05 = .4\hat{\pi}_1$$
$$\hat{\pi}_1 = -.125$$

Thus, $\hat{\pi}_1$ takes on a negative value—mathematically correct, but meaningless in terms of a solution to the problem at hand. This result (and similar unusual ones) can happen (still assuming truthful responses) whenever the actual results of the randomizing device vary to some (generally, large) degree from the expected results, i. e., when the *actual* P_1 value is, by chance alone, significantly different from the *expected* P_1. (This occurrence is relatively unlikely, though, for even moderately large sample sizes.)

4. The subscripts 1 and 2 on the *P*-values denote the randomizing probabilities for Samples No. 1 and 2, respectively. Note that P_1 must not be set the same as P_2, i.e., $P_1 \neq P_2$. This is a necessary mathematical condition for solving simultaneous equations.

REFERENCES

1. Campbell, Cathy, and Joiner, Brian L., "How to Get the Answer without Being Sure You've Asked the Question," *The American Statistician*, 27: 229-231, 1973.
2. Cochran, W.G., *Sampling Techniques* (2nd ed.), John Wiley and Sons, New York, 1953.
3. "Estimation of Birth Rates and Population Change from Sample Surveys; Progress Reports Nos. 1-5," *Project No. SU-221*, Research Triangle Institute, Durham, N.C., 1965-1966.

4. Fox, David J., *The Research Process in Education.* Holt, Rinehart and Winston, New York, 1969.
5. Greenberg, Bernard G.; Abul-Ela, Abdel-Latif A.; Simmons, Walt R.; and Horvitz, Daniel G., "The Unrelated Question Randomized Response Model: Theoretical Framework," *Journal of the American Statistical Association,* 64: 520-539, 1969.
6. Greenberg, Bernard G.; Kuebler, Roy R., Jr.; Abernathy, James R.; and Horvitz, Daniel G., "Application of the Randomized Response Technique in Obtaining Quantitative Data," *Journal of the American Statistical Association,* 66: 243-250, 1971.
7. Hansen, M.H.; Hurwitz, W.N.; and Madow, W.G., *Sampling Survey Methods and Theory,* Vol. II, John Wiley and Sons, New York, 1953.
8. Hansen, M.H.; Hurwitz, William N.; Marks, Eli S.; and Mauldin, W. Parker, "Response Errors in Surveys," *Journal of the American Statistical Association,* 46: 147-190, 1951.
9. Hansen, M.H.; and Marks, E.S., "Influence of the Interviewer on the Accuracy of Survey Results," *Journal of the American Statistical Association,* 53: 635-655, 1958.
10. Horvitz, D.G.; Shah, B.V.; and Simmons, Walt R., "The Unrelated Question Randomized Response Model," *Social Statistics Section: Proceedings of the American Statistical Association* (1967), 65-72.
11. Selltiz, Claire; Jahoda, Marie; Deutsch, Morton; and Cook, Stuart W., *Research Methods in Social Relations* (Revised Edition), Holt, Rinehart and Winston, New York, 1964.
12. Stephan, F.F.; and McCarthy, P.J., *Sampling Opinions,* John Wiley and Sons, New York, 1963.
13. Warner, Stanley L., "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias," *Journal of the American Statistical Association,* 60: 63-69, 1965.
14. Warner, Stanley L., "The Linear Randomized Response Model," *Journal of the American Statistical Association,* 66: 884-888, 1971.

COMPREHENSION BY COLLEGE STUDENTS OF TIME-COMPRESSED LECTURES

LORETTA ADELSON
Brooklyn College

ABSTRACT

Time-compressed speech has been suggested as an educational medium for sighted and blind students, with researchers reporting high comprehension of short, unrelated passages. For this study a repeated measurement design was used in presenting equated one-hour lectures to 200 students at 175 and 275 words per minute. Mean comprehension scores, standard deviations, t-tests and percentiles were used to analyze the data. Standard deviations were stable, and time-compressed comprehension scores showed losses that were significant at the .01 level. The Fairbanks efficiency index is thus questioned, and it is further suggested that time-compressed speech as an educational medium needs further study.

TIME-COMPRESSED SPEECH refers to recorded speech which has been altered in time. In 1952, Fairbanks, Everitt, and Jaeger produced an electronic device which was capable of picking up millimeters of sound at a predetermined rate, discarding those sounds, and then abutting the remaining sounds. The result was speech recorded at any desired rate without the loss of any one complete phoneme, continuity, or the original pitch and vocal quality (2).

In order to test comprehension of time-compressed speech, Fairbanks, Guttman, and Miron composed two factual passages, each over 1500 words in length. They reported that at 282 wpm, "the response was slightly less than 90% "of the response at 141 wpm (5:18). Encouraged

by such results, time-compressed speech was then given serious consideration as a useful medium in the education of the blind as well as the sighted, and research in its use was supported by the U. S. Office of Education (9, 10).

On the whole, materials that have been used have been comparatively short in length. One researcher used passages that ran for 18 to 20 minutes at the normal rate, 175 wpm (10). More frequently, however, researchers have relied upon standardized listening comprehension tests (11, 12, 14, 15). The use of such tests may be questioned since the tests are comprised of unrelated passages varying in length from 25 seconds to 3 minutes and 45 seconds. Each passage is followed immediately by a written test of comprehension. In a critical evaluation of the STEP Listening Test Lorge wrote, "In terms of the objectives stated by the teachers and educators, this makes for relatively short listening comprehension situations. Among high school and college students the expectation should be for significantly longer listening time and for more questions" (13:572). The customary college lecture hour would appear to be a more realistic condition to use in the assessment of the listening comprehension of time-compressed speech.

The purpose of this study was to examine the degree of comprehension that was shown by a group of college students when listening to a one-hour lecture of educational materials at the normal rate of speech, 175 wpm, as compared with the degree of comprehension that was shown by the same group of college students when listening to an equated one-hour lecture at a time-compressed rate of 275 wpm for 40 minutes.[1]

Method

Subjects

Two hundred undergraduate students at Brooklyn College of the City University of New York volunteered for this study. All spoke English as their first language. Only those students who wrote the correct answers to all 21 questions of the Harvard Psycho-Acoustic Laboratory Auditory Test No. 12 participated (3:490).

Materials

Six passages from English history, equated by Friedman and Orr, were used to comprise two one-hour lectures.[2] The passages had been equated for length, the average number of words per sentence, the average number of syllables per 100 words, listening grade as determined by Rogers' listenability formula, reading ease measured by the Flesch formula, the number of independent clauses, and the average number of words not on the Dale list. Three passages were recorded on one tape and constituted Lecture A. Three other passages were recorded on another tape and constituted Lecture B. Both lectures were prepared at the normal and time-compressed rates in order to satisfy the design as shown in Figure 1. In order to familiarize the students with time-compressed speech, a two-minute selection from a seventh passage was taped so as to immediately precede the time-compressed lecture. The lecture ran for 60 minutes at the normal rate, and for 40 minutes at the time-compressed rate.

Listening comprehension tests for the six passages had been devised by Friedman and Orr. The test for each lecture had, in effect, 75 five-option multiple choice items. Scores were corrected for chance and given to the nearest whole number. Comprehension tests were administered to 34 students without presentation of the lectures. The mean comprehension scores for Lectures A and B were 6.44 and 3.71, respectively, and the average of these two mean scores is 5.08. Raw scores only were used in the study.

Listening materials were recorded on magnetic tape supplied by the American Institutes for Research. They were presented free-field by the use of a Wollensak Model 1520 at 3¾ speed in a semi-soundproofed room (16, 18). Intelligibility of the lectures was considered to be good.

Design

This was a repeated measurement study in which subjects were compared with their own performances. The design used controlled for any differences in Lecture A and Lecture B difficulty levels (Figure 1).

SUBJECTS	SESSION 1	SESSION 2
I (n = 50)	Lecture A - Normal Followed by Test A	Lecture B - Time-Comp. Followed by Test B
II (n = 50)	Lecture B - Time-Comp. Followed by Test B	Lecture A - Normal Followed by Test A
III (n = 50)	Lecture B - Normal Followed by Test B	Lecture A - Time-Comp. Followed by Test A
IV (n = 50)	Lecture A - Time-Comp. Followed by Test A	Lecture B - Normal Followed by Test B

Figure 1.—Experimental design used to control lecture A and Lecture B difficulty levels

Procedure

After having performed a pilot study with 20 students, 200 students were assigned consecutively and at random to Groups I, II, III, and IV. To encourage a high level of performance, students were told that those who achieved a score in the upper 10 percent would receive a letter of commendation.

Table 1.—Mean Comprehension Scores and Standard Deviations for Groups I, II, III, IV, and the Total Sample

Group	Rate	Passage 1		Passage 2		Passage 3		Total Lecture	
		Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.
I	Normal	9.18	4.59	9.46	4.70	8.28	5.73	27.14	13.22
N = 50	Time-Comp.	5.04	4.51	4.88	3.53	7.18	4.02	17.10	10.30
II	Normal	9.50	5.59	8.90	5.16	7.06	6.14	25.46	15.30
N = 50	Time-Comp.	6.34	4.93	4.88	4.13	7.32	5.84	18.54	13.02
III	Normal	7.84	4.82	6.30	4.36	9.20	5.65	23.34	12.35
N = 50	Time-Comp.	6.64	3.83	5.18	3.24	5.14	4.35	16.82	8.82
IV	Normal	8.32	4.79	5.36	4.25	9.94	5.65	23.62	12.78
N = 50	Time-Comp.	6.32	4.24	4.48	4.00	3.38	3.44	14.18	9.72
Total	Normal	8.71	4.97	7.51	4.91	8.62	5.85	24.89	13.45
N= 200	Time-Comp.	6.09	4.41	4.86	3.72	5.76	4.75	16.66	10.62

Raw scores were used for all computations. A comparison of comprehension was effected by obtaining the mean scores and standard deviations under both conditions for the total sample and its four component groups, and carrying out *t*-tests in order to determine whether there were any significant effects between: the normal and time-compressed rate; the first and second order of presentation of the lectures; the form of Lecture A and the form of Lecture B. To determine the significance of the differences between the normal rate and time-compressed rate mean comprehension scores, for the total sample and all subgroups, *t*-tests were also used. Percentiles were computed manually for the normal rate scores. These were used as a reference normative base group for the time-compressed scores in order to observe the degree of separation between the two conditions.

Discussion

Mean Comprehension Scores and Standard Deviations

The mean scores and standard deviations for the total sample and the four component groups are given by passage and total lecture at both rates in Table 1. Of prime interest, the 200 subjects had a mean score of 24.89 at the normal rate, and a mean score of 16.66 at the time-compressed rate, showing a decrement of 8.23 points, or approximately one-third of the normal rate score.

The standard deviations may be considered to be uniform from passage to passage and from condition to condition. Differences of the standard deviations given for Groups I, II, III, and IV are interpreted as reflecting the differences between the groups due to random sampling rather than having used matched groups. The larger differences between the standard deviations of the total scores and the passages reflects the increased length of the materials and the resulting wider range of scores.

t-tests

In order to determine if there were significant differences associated with the order of presentation (Session 1 versus Session 2), the form of the lecture (Lecture A versus Lecture B), and the rate (Normal versus Time-Compressed), *t*-tests were computed for the pairs of conditions. Tables 2 and 3 indicate that at the .05 level, for a two-tail test, there was no significant effect of order or form on the mean comprehension scores for all four component groups.

However, the effect of rate was statistically significant at the .01 level for three of the four conditions shown in Table 4. All four differences were in the same positive direction. The differences in size of the *t*-values for Lecture A under both conditions, as compared with the *t*-values for Lecture B under both conditions, suggest that rate produced a stronger effect on Lecture A than it did on Lecture B. This possibility is supported by the fact that as originally devised by Friedman and Orr, Lecture A contained ten questions more than Lecture B. This was adjusted in scoring. However, it is plausible that the use of a larger number of questions was due to a larger number of ideas per listening time. It is well to note that greater density of ideas in Lecture A might have interfered with apprehension of the materials at the time-compressed rate, but had no effect at the normal rate when there was adequate time for the processing of the concepts presented.

Also confirmed by *t*-tests was that the differences of the two comprehension scores were significant at the .01 level for the total sample and three of the four component groups (Table 5). The other component was significant at the .05 level. Differences are large and in the same positive direction.

Table 2.—Mean Comprehension Scores and Significance of Differences when Lecture is Presented First Versus Second

Group	Variable	Mean	S. D.	t
I	A_{1N}	27.14	13.22	0.59 N.S.*
II	A_{2N}	25.46	15.30	
IV	A_{1C}	14.18	9.72	-1.42 N.S.*
III	A_{2C}	16.82	8.82	
III	B_{1N}	23.34	12.35	-0.11 N.S.*
IV	B_{2N}	23.62	12.78	
II	B_{1C}	18.54	13.02	-0.61 N.S.*
I	B_{2C}	17.10	10.30	

A = Lecture A
B = Lecture B
C = Time-Comp Rate
N = Normal Rate
1 = Presented First
2 = Presented Second
*Not significant at .05 level for two-tail test

Table 3.—Mean Comprehension Scores and Significance of Differences when Lecture A Versus Lecture B is Presented

Group	Variable	Mean	S. D.	t
III	B_{1N}	23.34	12.35	1.48 N.S.*
I	A_{1N}	27.14	13.22	
IV	A_{1C}	14.18	9.72	-1.90 N.S.*
II	B_{1C}	18.54	13.02	
III	A_{2C}	16.82	8.82	-0.15 N.S.*
I	B_{2C}	17.10	10.30	
II	A_{2N}	25.46	15.30	0.65 N.S.*
IV	B_{2N}	23.62	12.78	

A = Lecture A
B = Lecture B
C = Time-Comp Rate
N = Normal Rate
1 = Presented First
2 = Presented Second
*Not significant at .05 level for two-tail test

Table 4.—Mean Comprehension Scores and Significance of Differences when Normal Rate Versus Time-Compressed Rate is Presented

Group	Variable	Mean	S. D.	t
I	A_{1N}	27.14	13.22	5.58*
IV	A_{1C}	14.18	9.72	
II	A_{2N}	25.46	15.30	3.46*
III	A_{2C}	16.82	8.82	
III	B_{1N}	23.34	12.35	1.89
II	B_{1C}	18.54	13.02	
IV	B_{2N}	23.62	12.78	2.81*
I	B_{2C}	17.10	10.30	

A = Lecture A
B = Lecture B
C = Time-Comp Rate
N = Normal Rate
1 = Presented First
2 = Presented Second
*Significant at .01 level for two-tail test

Table 5.–Mean Comprehension Scores for Normal and Time-Compressed Rates and Significance of Differences for All Groups

Group	Normal Rate Mean	Normal Rate S.D.	Time-Comp Rate Mean	Time-Comp Rate S.D.	t
All Subjects (N = 200)	24.89	13.45	16.66	10.62	6.80*
I (N = 50)	27.14	13.22	17.10	10.30	4.24*
II (N = 50)	25.46	15.30	18.54	13.02	2.45**
III (N = 50)	23.34	12.35	16.82	8.82	3.05*
IV (N = 50)	23.62	12.78	14.18	9.72	4.16*

*Significant at .01 level for a two-tail test

**Significant at .05 level for a two-tail test

Table 6.–Degree of Separation between Normal and Time-Compressed Quartiles

Rate	P_{25}	P_{50}	P_{75}
Normal	13.95	23.50	34.59
Time-Comp.	8.78	14.63	23.88

Table 7.–Percentiles for the Mean Comprehension Scores of All Subjects and the Four Component Groups, Showing Percentile Difference between the Normal and Time-Compressed Conditions:

Group	Normal Rate Mean	Percentile	Time-Comp Rate Mean	Percentile*	Percentile* Difference
All Subjects N = 200	24.89	55	16.66	33	22
I N = 50	27.14	58	17.10	33	25
II N = 50	25.46	56	18.54	37	19
III N = 50	23.34	53	16.82	33	20
IV N = 50	23.62	54	14.18	25	29

*Normal rate scores used as reference normative base group.

Percentiles

The use of percentiles afforded a clear view of the magnitude of the separation between the two mean scores. Percentiles computed for the normal rate scores were used as the reference normative base group. Table 6 indicates that there was a difference between the two conditions that was almost equal to the semi-interquartile range. Thus, under time-compressed conditions, only about a quarter of the students did as well as the average student at the normal rate.

Table 7 indicates that the normal rate mean comprehension score for the total number of subjects was 24.89. Scores were positively skewed at the normal rate, so that this score was at the 55th percentile. The time-compressed mean comprehension score was 16.66. Scores were negatively skewed. This latter score fell at the 33rd percentile of the normal rate percentiles. Similar patterns were shown for all component groups.

The foregoing divergent scores should be kept in mind when reviewing the degrees of comprehension established by previous studies. Fairbanks, Guttman, and Miron recorded 90 percent as great comprehension for two passages that ran a little over five minutes each at a rate of 282 wpm (5:18). Orr and Friedman reported, "The main result of the experiment indicated that naive (untrained) subjects suffer relatively little loss (0–20%) in listening comprehension at speeds up to twice normal speaking rates (up to 325 wpm)" (16:ii). Foulke *et al.* reported no significant loss in comprehension of a scientific selection presented to 291 blind sixth, seventh, and eighth grade pupils at 275 wpm (7). The selection ran for seven and one-half minutes and was at the fifth to sixth grade level of readability. McCracken found no difference in comprehension when she presented four short selections at 160 wpm and another four short selections at 320 wpm from the Diagnostic Reading Tests, Section II (14). The entire session, including testing, lasted approximately 30 minutes. It is apparent that when a realistic length of listening material was presented to sighted college students, considerably less was comprehended at the normal rate. The length of uninterrupted listening time appears to be a critical factor in its influence upon time-compressed listening comprehension.

In light of the data given in this study, it would appear necessary to examine the efficiency index presented by Fairbanks (5) and used by Foulke *et al.* (7), Jester and Travers (12), and Sticht (18). Foulke and Sticht have expressed Fairbanks' efficiency index as a ratio between the comprehension score and the time used for presentation. They "found that learning efficiency increased as word rate was increased until a word rate of approximately 280 wpm was reached" (9:12). Foulke *et al.* concluded their study with the statement, "It was felt that those losses in comprehension that were statistically significant were not all educationally important, especially when the time saved in presenting the material was considered" (7:141). In view of the large loss in comprehension found in this study, it is important to point out that the efficiency index appears to rest on an underlying assumption that all items are equally easy or difficult to learn, and that all items are equally important or unimportant to their educational objectives. Nor does it take into consideration the density of the ideas presented. In effect, as the efficiency index now operates, it tends to put a stamp of approval on those students who learned less, simply because they spent less time doing so.

Conclusions

This study has assessed the comprehension by 200 Brooklyn College students of a one-hour lecture at 175 wpm as compared with their comprehension of an equated time-compressed lecture at 275 wpm. The one-hour lecture period was chosen because it was considered to be more realistically representative of the college students' listening experiences than the considerably shorter periods that have been used by researchers to date. A *t*-test established that the difference between these two comprehension scores was statistically significant. Further *t*-tests for three criteria employed in this study—order of presentation, form of lecture used, and rate used—proved rate alone to be statistically significant. Percentiles constructed using the normal rate comprehension scores as a reference normative base group showed that normal rate scores were positively skewed and that time-compressed rate scores were negatively skewed. This large drop in comprehension differs with findings reported by previous researchers.

The following conclusions are drawn:

1. The length of the stimulus materials used appears to be a critical factor. Time-compressed materials suffered a proportionately larger loss of comprehension than did the normal rate materials when an educationally realistic length of materials was used.
2. The efficiency index suggested by Fairbanks and used by subsequent researchers has been questioned operationally because it fails to take the following factors into consideration: the density of the ideas present in the listening materials; the number of items not learned; the importance of the items learned and not learned; the relative difficulty of the items learned and not learned; a criterion of acceptable comprehension stated in advance. It is suggested that an effective evaluation of an educational medium must take these factors into consideration.
3. The introduction of the efficiency index to justify the use of time-compressed speech as an educational medium has been further questioned because it puts a premium on learning less, provided that less time has been spent doing so; it shifts the educational emphasis from "how to educate more adequately" to "how to educate more groups"; and it encourages skimming rather than probing and analyzing.
4. It is suggested that time-compressed speech might be of educational value when used to develop controlled, graduated rates of spoken practice materials for developing speed in stenography and stenotyping.

5. Time-compressed speech may also serve a useful purpose in vigilance tasks which demand a choice among a limited number of alternatives, thus making the task one of discrimination or intelligibility rather than comprehension.

6. Further studies that would concentrate on the establishment of what might be considered to be an appropriate length of stimulus materials for a realistic testing situation at different educational levels would be valuable.

7. Further studies that would contribute toward the establishment of criteria for the measurement of idea density within test materials would help to make the interpretation of comprehension scores more meaningful.

FOOTNOTES

1. John B. Carroll (1:60) has suggested that syllables per minute is a better measure of rate. As noted under *Materials*, the number of syllables per 100 words was one of the criteria used in the equating of the lectures used for this study.

2. The author is indebted to David B. Orr and Herbert L. Friedman of the American Institutes for Research, Silver Spring, Maryland, who granted the use of their materials.

REFERENCES

1. Carroll, John B., *Language and Thought,* Prentice-Hall, Englewood, N.J., 1964.
2. Cramer, H. Leslie, "An Introduction to Speech Compression Techniques: The Early Development of Speech Time Compression Concept and Technology," *Proceedings of the Second Louisville Conference on Rate and/or Frequency-Controlled Speech,* University of Louisville, Ky., 1971, 6-20.
3. Davis, Hallowell and Silverman, S. Richard, *Hearing and Deafness* (3rd ed.), Holt, Rinehart and Winston, New York, 1970.
4. Duker, Sam, *Time-Compressed Anthology and Annotated Bibliography* (3 vols.), Scarecrow Press, Metuchen, N.J., 1974.
5. Fairbanks, Grant; Guttman, Newman; and Miron, Murray S., "Effects of Time Compression upon the Comprehension of Connected Speech," *Journal of Speech and Hearing Disorders,* 22:10-19, 1957.
6. Foulke, Emerson, "Comparison of Comprehension of Two Forms of Compressed Speech," *Exceptional Children,* 33:169-173, 1966.
7. Foulke, Emerson; Amster, Clarence H.; Nolan, Carson Y.; and Bixler, Ray H., "The Comprehension of Rapid Speech by the Blind," *Exceptional Children,* 29: 134-141, 1962.
8. Foulke, Emerson; and Sticht, Thomas, "A Review of Research on Time-Compressed Speech," *Proceedings of the Louisville Conference on Time-Compressed Speech,* University of Louisville, Louisville, Ky., 1967, 69-76.
9. Foulke, Emerson; and Sticht, Thomas, "A Review of Research on the Intelligibility and Comprehension of Accelerated Speech," in Foulke, E. (ed.), *The Comprehension of Rapid Speech by the Blind: Part III,* Non-Visual Perceptual Systems Laboratory, University of Louisville, Louisville, Ky., 1969.
10. Friedman, Herbert L.; Orr, David B.; and Graae, Cynthia N., *Further Research on Speeded Speech as an Educational Medium,* Final report, Parts 1-5, American Institutes for Research in Behavioral Sciences, Silver Spring, Md., July 1965-September 1967. (Project No. 5-0801, Grant No. OE 7-48-7670--267. U.S. Department of Health, Education and Welfare, Office of Education).
11. Gore, George V., "A Comparison of Two Methods of Speeded Speech Presented to Blind Senior High School Students," unpublished Ed.D. dissertation, Teachers College, Columbia University, 1968.
12. Jester, Robert E.; and Travers, R.M.W., "Comprehension of Connected Meaningful Discourse as a Function of Rate and Mode of Presentation," *Journal of Educational Research,* 59: 297-302, 1966.
13. Lorge, Irving, Article Evaluating the Sequential Tests of Educational Progress, in Buros, O.K. (ed.), *The Fifth Mental Measurements Yearbook.* The Gryphon Press, Highland Park, N.J., 1959, 579-582.
14. McCracken, Sally, "Comprehension for Immediate Recall of Time Compressed Speech as a Function of Sex and Level of Activation," *Proceedings of the Second Louisville Conference on Rate and/or Frequency-Controlled Speech,* University of Louisville, Louisville, Ky., 1971, 313-319.
15. Michel, Leslie A., "Levels of Comprehension and Activation under Conditions of Time-Compressed Speech," unpublished Master's Thesis, Florida State University, June 1970.
16. Orr, David B.; and Friedman, Herbert L., *Research on Speeded Speech as an Educational Medium,* Progress Report, Grant No. 7-48-7670-203, U.S. Department of Health, Education and Welfare, Office of Education, Washington, D.C., June 1964.
17. Sticht, Thomas, "Failure to Increase Learning Using the Time Saved by the Time Compression of Speech," *Journal of Educational Psychology,* 62: 55-59, 1971.
18. Sticht, Thomas, "Some Relationships of Mental Aptitude, Reading Ability, and Listening Ability Using Normal and Time-Compressed Speech," *Journal of Communication,* 18: 243-258, 1968.

U. S. POSTAL SERVICE

STATEMENT OF OWNERSHIP, MANAGEMENT AND CIRCULATION

(Act of August 12, 1970: Section 3685, Title 39, United States Code

1. TITLE OF PUBLICATION
The JOURNAL OF EXPERIMENTAL EDUCATION

2. DATE OF FILING
9/30/75

3. FREQUENCY OF ISSUE
Quarterly

3A. ANNUAL SUBSCRIPTION PRICE
$12.50

4. LOCATION OF KNOWN OFFICE OF PUBLICATION *(Street, city, county, state and ZIP code) (Not printers)*
4000 Albemarle St., N.W. Suite 302; Washington, DC 20016

5. LOCATION OF THE HEADQUARTERS OR GENERAL BUSINESS OFFICES OF THE PUBLISHERS *(Not printers)*
Same as above

6. NAMES AND ADDRESSES OF PUBLISHER, EDITOR, AND MANAGING EDITOR

PUBLISHER *(Name and address)* Helen Dwight Reid Educational Foundation
4000 Albemarle St., N.W.; Suite 302; Washington, DC 20016

EDITOR *(Name and address)*
John Schmid, Chairman, Dept. of Res. & Statistical Methodology, Univ. of Northern Colorado
Greeley, CO 80631

MANAGING EDITOR *(Name and address)*
Cornelius W. Vahle, Jr.; 4000 Albemarle St., N.W.; Suite 302; Washington, DC 20016

7. OWNER *(If owned by a corporation, its name and address must be stated and also immediately thereunder the names and addresses of stockholders owning or holding 1 percent or more of total amount of stock. If not owned by a corporation, the names and addresses of the individual owners must be given. If owned by a partnership or other unincorporated firm, its name and address, as well as that of each individual must be given.)*

NAME	ADDRESS
Helen Dwight Reid Educational Foundation	4000 Albemarle St., N.W. Suite 302 Washington, DC 20016

8. KNOWN BONDHOLDERS, MORTGAGEES, AND OTHER SECURITY HOLDERS OWNING OR HOLDING 1 PERCENT OR MORE OF TOTAL AMOUNT OF BONDS, MORTGAGES OR OTHER SECURITIES *(If there are none, so state)*

NAME	ADDRESS

9. FOR OPTIONAL COMPLETION BY PUBLISHERS MAILING AT THE REGULAR RATES *(Section 132.121, Postal Service Manual)*

39 U. S. C. 3626 provides in pertinent part: "No person who would have been entitled to mail matter under former section 4359 of this title shall mail such matter at the rates provided under this subsection unless he files annually with the Postal Service a written request for permission to mail matter at such rates."

In accordance with the provisions of this statute, I hereby request permission to mail the publication named in Item 1 at the reduced postage rates presently authorized by 39 U. S. C. 3626.

(Signature and title of editor, publisher, business manager, or owner)
Cornelius W. Vahle, Jr. Publisher

10. FOR COMPLETION BY NONPROFIT ORGANIZATIONS AUTHORIZED TO MAIL AT SPECIAL RATES *(Section 132.122 Postal Service Manual) (Check one)*

The purpose, function, and nonprofit status of this organization and the exempt status for Federal income tax purposes

☒ Have not changed during preceding 12 months ☐ Have changed during preceding 12 months *(If changed, publisher must submit explanation of change with this statement.)*

11. EXTENT AND NATURE OF CIRCULATION	AVERAGE NO. COPIES EACH ISSUE DURING PRECEDING 12 MONTHS	ACTUAL NUMBER OF COPIES OF SINGLE ISSUE PUBLISHED NEAREST TO FILING DATE
A. TOTAL NO. COPIES PRINTED *(Net Press Run)*	2,400	2,500
B. PAID CIRCULATION 1. SALES THROUGH DEALERS AND CARRIERS, STREET VENDORS AND COUNTER SALES	--	--
2. MAIL SUBSCRIPTIONS	2,000	2,000
C. TOTAL PAID CIRCULATION	2,000	2,000
D. FREE DISTRIBUTION BY MAIL, CARRIER OR OTHER MEANS SAMPLES, COMPLIMENTARY, AND OTHER FREE COPIES	100	100
E. TOTAL DISTRIBUTION *(Sum of C and D)*	2,100	2,100
F. COPIES NOT DISTRIBUTED 1. OFFICE USE, LEFT-OVER, UNACCOUNTED, SPOILED AFTER PRINTING	300	400
2. RETURNS FROM NEWS AGENTS	--	--
G. TOTAL *(Sum of E & F—should equal net press run shown in A)*	2,400	2,500

I certify that the statements made by me above are correct and complete.

SIGNATURE OF EDITOR, PUBLISHER, BUSINESS MANAGER, OR OWNER
Cornelius W. Vahle, Jr. Publisher

PS Form 3526 (Page 1)
Jan. 1975

THE CONTRIBUTION OF NONINSTRUCTIONAL ACTIVITIES TO COLLEGE CLASSROOM TEACHER EFFECTIVENESS

RONALD D. McCULLAGH
MELVIN R. ROY
Appalachian State University

ABSTRACT

The contribution of noninstructional activities to classroom teacher effectiveness was investigated by administering a questionnaire to 52 university faculty members. The criterion for classroom effectiveness was a student evaluation of each teacher. A simple correlation analysis and the multiple regression technique were used to evaluate the results of this study. A significant result of the study was the failure of noninstructional activities to have predictive value when student-perceived teacher effectiveness is used as the criterion. A second significant result was the negative effect that time spent in consulting had upon classroom teacher effectiveness. Implications for further research are discussed, including the suggestion that a reevaluation of the responsibility of the university to the community, and of education to society, may be necessary.

GOVERNMENT, INDUSTRY, and education are becoming increasingly dependent on American colleges and universities for current information and services, as well as personnel. That dependence has created increased pressure on faculty to perform duties such as consulting, research, and civic and academic committee work. College and university administrations encourage additional noninstructional activities, such as student counseling, professional affiliation, and publishing, for their academic value. Educational administration gives the following reasons for its encouragement, citing that noninstructional activities:

- –keep a professor abreast of his field;
- –result in the introduction of highly relevant material into classrooms;
- –stimulate a professor's desire to teach; and
- –introduce the process of systematic inquiry into the classroom (7).

Classroom teachers do not always share administration's positive regard for noninstructional activities. When conflict arises, teachers tend to rely on one of the following arguments, which contend that noninstructional activities:

- –leave too little time for classroom preparation;
- –result in the over-sophistication of materials presented in the classroom;
- –leave professors unavailable for conferences regarding matters pertaining to course work; and
- –make only limited contributions to the quality of teaching (7).

Each of the arguments against noninstructional activities is based on the assumption that instructional activities constitute the primary function of professors. It is that assumption which leads professors to question the contribution of noninstructional activities to college classroom teacher effectiveness.

The issue of what individual qualities most enhance teacher effectiveness remains largely unresolved. But attempts by education to identify influential qualities continue. Today the influence of administration on faculty to enlarge the instructional base by involvement in noninstructional activities is very real. In spite of administration insistence and faculty resistance in the matter, little, if any, evidence exists to prove or disprove the rationale for either case.

Verdicts of some of the most comprehensive studies made to date are inconclusive. Prior to the statement of the Committee on the Criteria of Teacher Effectiveness, Orville Brim (2) concluded that there were no consistent relations between teacher characteristics and effectiveness in teaching. In 1963, the *Handbook of Research on Teaching* (5)

reported that "teaching methods do not seem to make much difference" and that "there is hardly any evidence to favor one method over another." Furthermore, it reported, "until very recently, the approach to the analysis of teacher-pupil and pupil-pupil interaction . . . has tended to be unrewarding and sterile." After examining the data as well as the conclusions of nearly one hundred studies, Dubin and Taveggia (4) concluded that college teaching methods make no difference in student achievement as measured by final examinations on course content.

Little encouragement is offered by the massive report of *Equality of Educational Opportunity* (3). According to that report, when the social background and attitudes of individual students and their schoolmates are held constant, achievement is only slightly related to school characteristics.

Another very comprehensive study conducted by Stephens (9) supported Coleman's findings. Documenting his position with the educational variables of school attendance, instructional television, independent study, correspondence study, class size, individual consultation and tutoring, counseling, student concentration, student involvement, amount of time spent in study, job distraction, extracurricular activities, school size, supervisor-rated teacher quality, nongraded school, team teaching, ability grouping, progressivism vs. traditionalism, discussion vs. lecture, directive vs. nondirective teaching, variable testing, and programmed instruction, Stephens concluded that practically nothing seems to make a difference in the effectiveness of instruction.

Fortunately, subsequent studies give reason to question the pessimism. According to Bowles and Levin (1), the research design used in the Coleman study was "overwhelmingly biased in a direction that would dampen the importance of school characteristics." For example, expenditure was measured within an entire school district rather than within the given school in which the pupils were located. Hence, the expenditure-per-pupil was overstated for schools attended by lower-class students and understated for schools attended by students of higher social status. Faulty statistical models were also used, according to Bowles and Levin (1). The importance of a variable was measured by how much the proportion of variance in achievement explained was increased by adding that variable to the predictors. This procedure was followed without reference to the order in which the variables were added into the regression equation, and the order of the variables was such as to favor or overweight the family background characteristics. Despite the discrepancies in design, exclusive of family background characteristics, teacher characteristics accounted for higher variation than all other aspects of the school combined (3). In final reference to the Coleman study, Mood (8) has stated that "the present rudimentary state of our quantitative models does not permit us to disentangle the effects of home, school, and peers on student's achievement."

Two studies offer encouragement and direction to the pessimism that surrounds research in teacher effectiveness. Mood (8) advises that one way to improve research is to obtain better measures, of a larger number, of the teacher attributes that are significant to the ability of teachers to improve learning. He says that such measures will come closer to estimating the full effect of teachers, independent of home and school factors. The second suggestion is to aim these measures at process variables, "those human actions which transform the raw materials of input into opportunities for learning," (Gagne, 6), i.e., . . . teacher activities, rather than teacher characteristics. It is to the contribution of those teacher activities that this study is directed.

Method and Procedure

Following an administrative push for an increase in the quantity and quality of research at Appalachian State University, 1500 students and 52 faculty members participated in a study of the relevance of administrative policy. The purpose of this study was to determine if a relationship existed between college classroom teaching effectiveness and noninstructional activity. This resulted in an attempt to correlate student-perceived teacher effectiveness with self-reported noninstructional activity.

In order to assess noninstructional activity, an instrument was developed to measure the degree of faculty involvement in such noninstructional activities as research, professional affiliation, committee work, articles published, conferences and workshops sponsored or attended, books published, and professional consulting (Appendix A). After the faculty members had completed a noninstructional activity inventory, their students completed an inventory of teacher characteristics (Appendix B). That instrument was designed to elicit information on teacher effectiveness, the criterion variable.

The two instruments provided a total of 11 variables for each faculty member, from which followed a correlation analysis. This was intended to yield intercorrelations between noninstructional activity and teacher effectiveness. Finally, an attempt was made to analyze sets of predictor variables (noninstructional activities) in terms of their contributions to the prediction of the criterion, college classroom teacher effectiveness.

Ten noninstructional activity variables were used, as reported by the 52 sample faculty members. The variables are as follows:

1 - (ACS) - The number of academic committees served
2 - (MPO) - The number of memberships in professional organizations
3 - (AWC) - The number of academic workshops and conferences attended
4 - (AAP) - The number of academic articles published

5 - (ABP) - The number of academic books published
6 - (PCC) - The number of professional consulting contracts
7 - (HrACS)- The average hours per week spent in academic committee meetings or in preparation for academic committee meetings
8 - (HrPP) - The average hours per week spent in preparation of academic materials intended for publication
9 - (HrWC) - The average hours per week spent in academic workshops and conferences or in preparation for academic workshops and conferences
10 - (HrPCC) - The average hours per week spent in performance of professional consulting contracts

Variable 11 (TchEff) rated the general classroom teacher effectiveness, without reference to specific characteristics.

No coding, rating, grading or scaling was done on the first ten variables, since the reported data represented 100 percent of the data. An examination of Table 1 reveals the tabulations of means and standard deviations for all of the variables.

The average score was computed for student-perceived teacher effectiveness on each sample faculty member. The average was computed by dividing the cumulative scores on variable 11 by the total number of responses to that variable. For example, if five respondents scored a sample faculty member 3, 5, 7, 6, and 9 on that variable, his cumulative score on that variable would be 30. When his cumulative score was divided by the total number of responses, 30/5, the result was an average score of 6. On a scale of 1 to 9, high scores represented effectiveness in reference to that variable.

This study was designed to explore the contribution of noninstructional activity to college classroom teacher effectiveness. Although not specifically hypothesized, the issue of intercorrelations between noninstructional activity and classroom teacher characteristics was presented. The issue of intercorrelations among different noninstructional activities was also presented. Table 2 illustrates the intercorrelations of those data. The intercorrelation coefficients among the eleven variables are product-moment coefficients.

The most notable feature of the full correlation analysis was the lack of correlation. Variable 11 (overall teacher effectiveness) was used as the criterion variable, the measure of teacher effectiveness. Variable 10 measured the hours per week spent in performance of professional consulting contracts. As can be seen in Table 2, only variable 10 offered any appreciable correlation with teacher effectiveness. The significance of this will be discussed in the regression analysis.

The first ten of the eleven variables were items included in the noninstructional activity instrument that was completed by the 52 sample faculty members. Items 7, 8, 9, and 10 duplicated the subject matter of items 1, 2, 3, 4, 5, and 6. The purpose of this was to measure the data as done by university administrations, then to measure the same data in units deemed more appropriate by the writers. Without reference to the credibility of either unit of measure, Table 2 reveals low correlations between similar data measured in different units. Variables 1 and 7 (committee work), variables 3 and 9 (workshops and conferences), variables 4 and 8 (publications), variables 5 and 8 (publications), and variables 6 and 10 (consulting contracts) illustrate the low correlations, despite the similarity of data tested.

Table 1.—Means and Standard Deviations for All Variables

	Variable	Mean	Standard Deviation
1	ACS	1.90	1.29
2	MPO	2.79	1.36
3	AWS	2.56	2.66
4	AAP	1.10	2.11
5	ABP	0.04	0.19
6	PCC	0.81	1.37
7	HrACS	1.49	1.13
8	HrPP	4.40	5.81
9	HrWC	1.32	2.86
10	HrPCC	1.68	6.25
11	TchEff	7.29	0.96

Table 2.–Intercorrelations

Variable	1	2	3	4	5	6	7	8	9	10	11
ACS 1	00	13	27	05	09	25	55	-12	32	-05	09
MPO 2	13	00	41	-01	-04	09	00	-09	05	-04	28
AWC 3	27	41	00	10	03	46	16	-11	27	-06	26
AAP 4	05	-01	10	00	-06	11	00	46	00	-05	03
ABP 5	09	-04	03	-06	00	32	09	12	08	-02	06
PCC 6	25	09	46	11	32	00	36	03	19	06	16
HrACS 7	55	00	16	00	09	36	00	11	41	02	-11
HrPP 8	-12	-09	-11	46	12	03	11	00	-11	-11	-05
HrWC 9	32	05	27	00	08	19	41	-11	00	03	12
HrPCC 10	-05	-04	-06	-05	-02	06	02	-11	03	00	-32
TchEff 11	09	28	26	03	06	16	-11	-05	12	-32	00

In addition to the simple correlation analysis, the investigator used multiple linear regression to determine the unique contribution of proper sets of the predictor variables, 1-10, to the production of the criterion, classroom teacher effectiveness. The contribution of a set of variables to prediction was measured by the difference between two squares of multiple correlation coefficients (RSs), one obtained for a regression model in which all predictors are used, called the *full model* (FM), and the other obtained for a regression equation in which the proper subset of variables under consideration have been deleted, called the *restricted model* (RM). The difference between the two RSs was tested for statistical significance with the variance ratio test.

The unique contribution of a variable to the prediction of a criterion may be interpreted in several ways, one of which seemed most reasonable to the writers. If a variable was making a unique contribution, then knowledge of that variable furnished information about the criterion. If a variable was making a unique contribution, then the respondents to the questionnaire, who were unlike on the variable but who were exactly alike or were matched on the other predictors, would differ on the criterion.

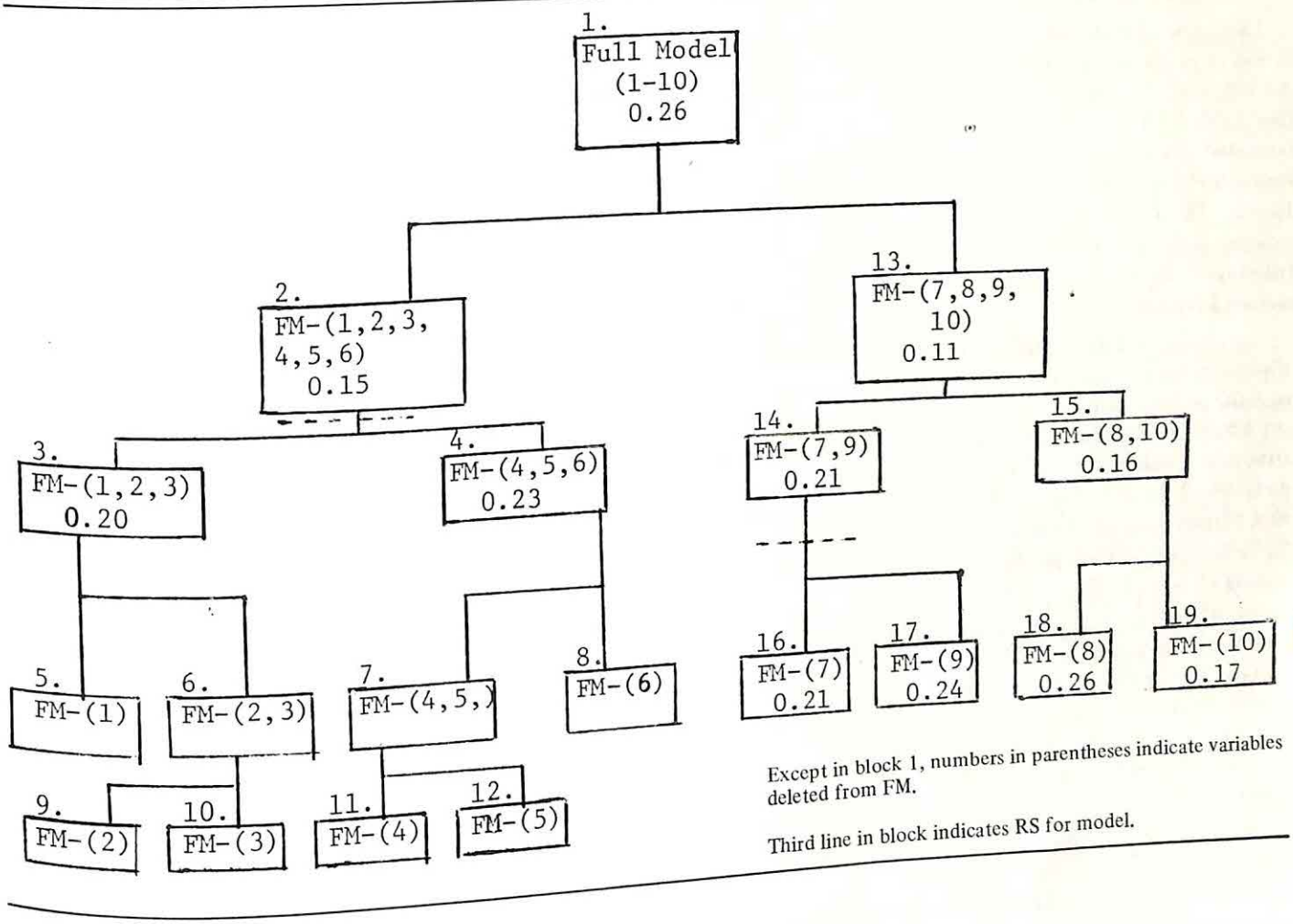
It was desirable to group predictors into logical sets, subsets, sub-subsets, and so forth down to the individual variables. The hierarchical grouping enabled the investigator to eliminate unnecessary tests. Subjective analysis of the predictors in this study suggested that they formed a hierarchical pattern as shown in Table 3.

A schematic was made to guide the sequence of tests; Table 4 illustrates the schematic. The topmost block, number 1, indicates that variables (1-10) were to be used as predictors in the FM. The next two sets, blocks 2 and 13, represent the numerically reported noninstructional activity and the hourly reported noninstructional activity.

Table 3.–Hierarchy of Variables

Numerically reported noninstructional activity:	
	(1, 2, 3, 4, 5, 6)
Nonremunerating involvement	(1, 2, 3)
Administrative	(1)
Professional	(2, 3)
Membership	(2)
Participation	(3)
Remunerating involvement	(4, 5, 6)
Publishing	(4, 5)
Articles	(4)
Books	(5)
Consulting	(6)
Hourly reported noninstructional activity	
	(7, 8, 9, 10)
Nonremunerating involvement	(7, 9)
Administrative	(7)
Professional	(9)
Remunerating involvement	(8, 10)
Publishing	(8)
Consulting	(10)

Table 4.–Schematic for Regression Models



Except in block 1, numbers in parentheses indicate variables deleted from FM.

Third line in block indicates RS for model.

The coding of the variables, for example, FM-(1, 2, 3, 4, 5, 6), as in block 2, indicates that the investigator wanted to obtain the RS, third line of the block, 0.15 in the present example, for a RM in which variables (1, 2, 3, 4, 5, 6) were eliminated from the FM. The RS for this RM could then be compared with the RS for the FM. A large difference or drop would indicate that the set (1, 2, 3, 4, 5, 6) was making a unique contribution to the prediction efficiency of the criterion. A small or zero drop would indicate that any relationship found for the variables (1, 2, 3, 4, 5, 6), collectively or individually, is not independent of variables (7, 8, 9, 10). If a drop for a set is not significant, then further tests of subsets of those variables are unnecessary, and an investigator may proceed to examine other lines of the hierarchy. The full proliferation of possible tests desired by the investigator is shown by Table 4. Significant models appear in the text of this article; those models beneath the termination line were significant and appear in Appendix C.

Discussion

Hypothesis 1 stated that there is no difference between the predictive efficiency of the full model and that of the zero model. (See Table 5.)

The regression of the criterion, teacher effectiveness, on all of the predictors, variables (1-10), had an RS equal to 0.26. The square root of this value, 0.51, was the correlation coefficient between the criterion and the best weighted composite of the predictors. This was substantially higher than the best individual predictor. Undoubtedly this RS represented an overfit of some unknown amount, an amount which would best be determined by cross validation with new data. Nevertheless, it seemed reasonable to conclude that whether a teacher will be effective can be predicted with a moderate degree of accuracy from knowledge of the noninstructional activity in which he participates.

Hypothesis 2 stated that there is no significant gain or loss in predictive efficiency when the numerically reported noninstructional activities are deleted from the full model. (See Table 6.) The restricted model, FM-(1, 2, 3, 4, 5, 6), which consisted of a model from which the numerically reported noninstructional activity had been deleted, had an RS equal to 0.15. The drop, 0.26 - 0.15 = 0.11, was statistically insignificant, indication that the entire set or subsets of numerically reported noninstructional activity were not making a unique contribution to the prediction of the criterion. Further analysis seemed pointless. In Table 4,

the dotted line following Model 2 illustrates the conclusion of the analysis of variables (1, 2, 3, 4, 5, 6).

Hypothesis 3 stated that there is no significant gain or loss in predictive efficiency when the hourly reported noninstructional activities are deleted from the full model. (See Table 7.) The restricted model, FM (7-10), which consisted of a model from which the hourly reported noninstructional activity had been deleted, had an RS equal to 0.11. The drop, 0.26 – 0.11 = 0.15, was statistically significant, indicating such activity was making a unique contribution to the prediction of the criterion. Further analysis seemed in order.

As shown in Table 8, the RS of 0.21 for the next subset, FM-(7, 9), which consisted of deleting the committee work variable and the conferences and workshops variable from the FM, indicated that these variables were making no contribution which could not be explained by the other eight variables of the FM. Thus, at this stage, it was concluded that further analysis of variables (7, 9) was unnecessary. In Table 4, the dotted line following Model 14 shows that testing of further subsets of this set was terminated at this point. Whatever relationship existed between the criterion and these two variables could be explained by other predictors. The investigator proceeded to test the collateral model, FM-(8, 10), Model 15, Table 9. The RS of 0.16 suggested that the variables, publications, and consulting contracts should be analyzed further. The model for the data had an RS of 0.16, which was significantly less than the 0.26 for the full model. It was concluded that further analysis of variables (8, 10) was in order.

Finally, when the variables (8, 10), publications and consulting contracts, were examined as shown in Table 10, the restricted model, Model 18, FM-(8), had an RS of 0.26. This amounted to no difference from the RS of 0.26 for the full model. It was concluded that publications made no contribution to the prediction of the criterion.

As shown in Table 11, Model 19, FM (10), consulting contracts, had an RS of 0.17, which was significantly less than that of the FM. Thus, it was concluded that of all the hourly reported noninstructional activity variables, only the average hours per week spent in performance of professional consulting contracts make a unique contribution to the prediction of the criterion, college classroom teacher effectiveness.

Conclusions

Therefore, the following conclusions were drawn in regard to the three specific hypotheses of the study:

Hypothesis 1, that there is no difference between the predictive efficiency of the full model and that of the zero model, was rejected. It seemed reasonable to conclude that whether a teacher will be effective can be predicted with a moderate degree of accuracy from knowledge of the noninstructional activity in which he participates.

Hypothesis 2, that there is no significant gain or loss in predictive efficiency when the numerically reported noninstructional activities are deleted from the full model, was accepted. It seemed reasonable to conclude that whether a teacher will be effective cannot be predicted with accuracy from knowledge of the numerically reported noninstructional activity in which he participates.

Hypothesis 3, that there is no significant gain or loss in predictive efficiency when the hourly reported noninstructional activities are deleted from the full model, was rejected. It seemed reasonable to conclude that whether a teacher will be effective can be predicted with a moderate degree of accuracy from knowledge of the hourly reported noninstructional activity in which he participates. Specifically, it was variable 10, hours per week spent in performance of consulting contracts which provided a unique contribution to prediction: increased involvement in consulting contracts having the effect of decreased college classroom teacher effectiveness.

The question of whether significant intercorrelations exist between noninstructional activities and college classroom teacher effectiveness was answered in the negative. Only the negative influence of consulting deviated from that general conclusion. The predictor variables were selected because they were most often recommended by college administrators and because they represented reasonably accessible types of information which might have some predictive relationship to teacher effectiveness. But the most notable feature of the data was their failure to correlate. The noninstructional activities had neither positive nor negative effects on teacher effectiveness. Even when the data were subjected to regression analysis in sets, the relationship between the noninstructional activities and teacher effectiveness was insignificant. As evidenced by the failure of data to correlate, and the failure of significant differences to appear between tested subsets, noninstructional activity made no contribution to classroom teacher effectiveness.

It would appear that the traditional predicting and evaluating standards of college classroom teacher effectiveness bear little or no relation to the subject. While a professor's noninstructional involvement may be an indicator of professional dedication, it is not an indicator of classroom effectiveness. If classroom instruction is to remain the primary function of professors, the traditional noninstructional activity criterion of teacher effectiveness must be discontinued, particularly for purposes of hiring and promoting of faculty.

The implications of these findings seem to offer two courses of action to college administrators. The first is for administration to search for new ways to help faculty increase classroom teacher effectiveness. The second is for administration to seek additional evidence in support of its claim of increased teacher effectiveness through involvement in noninstructional activity. These findings in no way

Table 5.–Regression Summary for Full Model Compared to the Zero Model

RS (FM) = 0.2663 RS (RM) = 0.0000	df (num) = 10 df (den) = 41	F = 1.28 PR = 0.1785

Table 6.–Regression Summary for Full Model Compared to Numerically Reported Noninstructional Activity

RS (FM) = 0.2663 RS (RM) = 0.1518	df (num) = 5 df (den) = 41	F = 1.28 PR = 0.2906

Table 7.–Regression Summary for Full Model Compared to Hourly Reported Noninstructional Activity

RS (FM) = 0.2663 RS (RM) = 0.116	df (num) = 4 df (den) = 41	F = 2.6 PR = 0.0896

Table 8.–Regression Summary for Full Model Compared to Hourly Reported Nonremunerating Involvement

RS (FM) = 0.2663 RS (RM) = 0.2124	df (num) = 2 df (den) = 41	F = 1.51 PR = 0.2321

Table 9.–Regression Summary for Full Model Compared to Hourly Reported Remunerating Involvement

RS (FM) = 0.2663 RS (RM) = 0.1694	df (num) = 2 df (den) = 41	F = 2.71 PR = 0.0769

Table 10.–Regression Summary for Full Model Compared to Hourly Reported Publishing

RS (FM) = 0.2663 RS (RM) - 0.2663	df (num) = 1 df (den) = 41	F = 0.00 PR = 0.9816

Table 11.–Regression Summary for Full Model Compared to Hourly Reported Consulting

RS (FM) = 0.2663 RS (RM) = 0.1705	df (num) = 1 df (den) = 41	F = 5.36 PR = 0.0243

suggest that noninstructional activities are without redeeming qualities. But a reevaluation of the responsibility of the university to the community, and of education to society, may be necessary for future justification of noninstructional activities in general.

Appendix A

ONLY YOU HAVE THIS INFORMATION, SO PLEASE HELP!

You were randomly selected as one of sixty Appalachian State University faculty members to participate in a teacher effectiveness study. This research represents an attempt to measure the contribution of non-teaching activities to overall classroom teacher effectiveness.

The purpose of this instrument is to determine the total amount of non-teaching activity in areas specified, in which ASU faculty members participated, during the period of December 1, 1971, to November 30, 1972.

It is *not* necessary to itemize the activities.

_____number of academic committees served

_____number of memberships in professional organizations

_____number of academic workshops and conferences attended

_____number of academic articles published

_____number of academic books published

_____number of professional consulting contracts

_____average hours per week spent in academic committee meetings or in preparation for academic committee meetings

_____average hours per week spent in preparation of academic materials intended for publication

_____average hours per week spent in academic workshops and conferences or in preparation for academic workshops and conferences

_____average hours per week spent in performance of professional consulting contracts

If this instrument is to serve the purpose for which it was intended, it must be followed by a student evaluation of you. Please indicate a class, hour, and room from which a sample evaluation may be drawn, prior to the conclusion of the current quarter.

_____class _____ hour_____ room number

Thank you for more help than one has a right to ask of you.

RETURN TO: Ron McCullagh
Business Administration
Campus

Appendix B

University of Northern Colorado
Bureau of Research
Professional Inventory

"Compared to other courses and other teachers, I would rate this course or this instructor–"

	Low Av. High
1. Course objectives are stated, followed, attained.	1 2 3 4 5 6 7 8 9
2. Assigned work is appropriate in amount and level.	1 2 3 4 5 6 7 8 9
3. The materials used (text, films, handouts, etc.) would rate	1 2 3 4 5 6 7 8 9
4. Everything considered, I would rate the worth of this course to me	1 2 3 4 5 6 7 8 9
5. The instructor's genuine interest in students	1 2 3 4 5 6 7 8 9
6. The instructor's communication skills–lecturing, questioning, answering, discussing	1 2 3 4 5 6 7 8 9
7. The instructor's professional qualities–thorough knowledge of the subject	1 2 3 4 5 6 7 8 9
8. The instructor's professional qualities–preparation for each class	1 2 3 4 5 6 7 8 9
9. The instructor makes difficult topics easy to understand	1 2 3 4 5 6 7 8 9
10. The instructor identifies what he considers important.	1 2 3 4 5 6 7 8 9
11. The instructor's personal characteristics–mannerisms and dress	1 2 3 4 5 6 7 8 9
12. The instructor's personal characteristics–is dynamic and energetic, enjoys teaching	1 2 3 4 5 6 7 8 9
13. The instructor's interpersonal relationships with students–fair, approachable, honest.	1 2 3 4 5 6 7 8 9
14. Ability to demonstrate skills and techniques	1 2 3 4 5 6 7 8 9
15. The instructor's interpersonal relationships with students–has a sense of humor	1 2 3 4 5 6 7 8 9
16. The instructor's availability and promptness–conferences and office hours	1 2 3 4 5 6 7 8 9
17. The instructor encourages and provides time for questions and discussion.	1 2 3 4 5 6 7 8 9
18. Everything considered, this instructor rates	1 2 3 4 5 6 7 8 9
19. (Special item may be used by instructor.)	1 2 3 4 5 6 7 8 9
20. (Special item may be used by instructor.)	1 2 3 4 5 6 7 8 9

Appendix C

Regression Summary For Full Model Compared To Numerically Reported Nonremunerating Involvement

RS (FM) = 0.2663 RS (RM) = 0.2028	df (num) = 4 df (den) = 41	F = 0.89 PR = 0.5180

Regression Summary For Full Model Compared To Numerically Reported Remunerating Involvement

RS (FM) = 0.2663 RS (RM) = 0.2394	df (num) = 4 df (den) = 41	F = 0.38 PR = 0.8257

Regression Summary For Full Model Compared To Numerically Reported Time Spent in Administrative Involvement

RS (FM) = 0.2663 RS (RM) = 0.2608	df (num) = 4 df (den) = 41	F = 0.31 PR = 0.5877

Regression Summary For Full Model Compared To Numerically Reported Time Spent in Professional Involvement

RS (FM) = 0.2663 RS (RM) = 0.2146	df (num) = 2 df (den) = 41	F = 1.45 PR = 0.2462

Regression Summary For Full Model Compared To Numerically Reported Publishing

RS (FM) = 0.2663 RS (RM) = 0.2662	df (num) = 2 df (den) = 41	F = 0.00 PR = 0.9967

Regression Summary For Full Model Compared To Numerically Reported Consulting

RS (FM) = 0.2663 RS (RM) = 0.2436	df (num) = 1 df (den) = 41	F = 1.27 PR = 0.2652

Regression Summary For Full Model Compared To Numerically Reported Time Spent in Professional Managership

RS (FM) = 0.2663 RS (RM) = 0.2277	df (num) = 1 df (den) = 41	F = 2.16 PR = 0.1459

Regression Summary For Full Model Compared To Numerically Reported Published Articles

RS (FM) = 0.2663 RS (RM) = 0.2663	df (num) = 1 df (den) = 41	F = 0.00 PR = 0.9614

Regression Summary For Full Model Compared To Numerically Reported Published Books

RS (FM) = 0.2663 RS (RM) = 0.2663	df (num) = 1 df (den) = 41	F = 0.00 PR = 0.9583

Regression Summary For Full Model Compared To Numerically Reported Administrative Involvement

RS (FM) = 0.2663 RS (RM) = 0.2179	df (num) = 1 df (den) = 41	F = 2.70 PR = 0.1040

Regression Summary For Full Model Compared To Numerically Reported Professional Involvement

RS (FM) = 0.2663 RS (RM) = 0.2458	df (num) = 1 df (den) = 41	F = 1.15 PR = 0.2901

Regression Summary For Full Model Compared To Numerically Reported Professional Managership

RS (FM) = 0.2663 RS (RM) = 0.2665	df (num) = 1 df (den) = 41	F = 2.16 PR = 0.1459

REFERENCES

1. Bowles, S. and Levin, H. M., "The Determinants of Scholastic Achievement–An Appraisal of Some Recent Evidence," *Journal of Human Resources,* 1968.
2. Brim, O. G., *Sociology and the field of education,* Russell Sage Foundation, New York, 1958.
3. Coleman, J. S., *Equality of Educational Opportunity,* U.S. Government Printing Office, Washington, D. C., 1966.
4. Dubin, R. and Taveggia, T., *The Teaching-Learning Paradox,* Center for Advanced Study of Educational Administration, University of Oregon, Eugene, 1968.
5. Gage, N. L., *Handbook of Research on Teaching,* American Educational Research Association, Rand McNally, Chicago, 1963.
6. Gagne, R. M., "How Can Centers for Educational Research Influence School Practices?", *Organizations for Research and Development in Education,* Phi Delta Kappan, Bloomington, Indiana, 1966.
7. Letchworth, G. and Hayes, L., "Effect of Faculty Research Activity on Instruction," *Oklahoma State Regents for Higher Education,* 1970.
8. Mood, A., *Do Teachers Make a Difference? A Report on Recent Research on Pupil Achievement,* Bureau of Educational Personnel Development, Office of Education, Washington, D.C., 1970.
9. Stephens, J. M., *The Process of Schooling,* Holt, Rinehart and Winston, New York, 1967.

DIRECTIONS FOR J.E.E. CONTRIBUTORS

The Journal of Experimental Education publishes specialized or technical education studies, treatises about the mathematics or methodology of behavioral research, and monographs of major current research interest.

ABSTRACT REQUIRED

An abstract of not more than 120 words must accompany each manuscript. It should precede the text, be designated *ABSTRACT,* and conform to the following standards:

1. In a research paper, include statements of (a) the problem, (b) the method, (c) the data, and (d) the conclusions.

2. In a review or discussion article, state the topics covered and the central thesis.

3. Standard abbreviations may be used freely if the meaning is clear. Use complete sentences and do not repeat information which is in the title.

TEXT

Usually research articles should have a clearly organized order of presentation. The following elements and their order is a guide and is not intended to be prescriptive.

The Problem. The nature, scope, and significance of the problem should be presented.

Related Research. Presentation and discussion of related research should be selective and should include references essential to clarify the problem under examination.

Methodology. This section should consist of hypotheses, description of the sample and sampling procedures, discussion of the variables, description of the data-gathering instruments (if well-known, their description may be omitted), and general description of the statistical procedures.

Presentation and Analysis of Data. Analysis of the data and conclusions about the hypotheses should be more than mere presentation. Rival hypotheses and significance for related studies and educational theory should be considered. Summary data (means, standard deviations, frequencies, correlation coefficients, etc.) should be presented in tabular or graphic form. Discussion of these data should be interpretive.

Summarizing Statements. A summary of conclusions and implications for education may supplement the abstract.

STYLE

Certain suggestions are offered here to achieve uniformity in publication style and to conserve the time and energy of the author and editorial staff in the preparation of copy. *A Manual of Style,* 12th ed., University of Chicago Press, Chicago, 1960, may be used as a style manual in preparation of manuscripts.

Two Copies Required. Copies should be double spaced with wide margins. Typed copies are preferred, but dittoed or mimeographed copies will be accepted if they are legible.

Subheads. Articles are frequently improved by the judicious use of subheads. However, avoid use of the title, INTRODUCTION, for a lead section.

Title. Try to use a short title, preferably no more than ten words. Avoid superfluous phrases, such as "A Comparison of . . .," "A Study of . . .," and "The Effectiveness of . . . ."

Tables. Prepare tables precisely as they are to appear in *The Journal,* caption each with a brief and to-the-point title, and number consecutively with arabic numerals: *Table 3.* INTERCORRELATION MATRIX, VARIABLES OPTIMALLY ORDERED.

Figures. Draw any graphs and charts in black ink on good quality paper, title each, and number consecutively, thus: *Figure 4.* SCHOOL ENROLLMENT. Send the original copy. We cannot accept mimeographed figures or charts. Data should not be framed, and ordinates and abscissas should be shortened to occupy as little space as possible.

Tables and Figures. Tables and figures must be original copies acceptable for reproduction. A charge will be assessed for any redrawing or re-typing of tables or figures.

Technical Symbols. All symbols, equations, and formulas must be clearly represented. Differentiate between numbers and letters (zero and the letter o; one and the letter l, etc.) Identify hand-drawn Greek letters in the margin. Align subscripts carefully.

Footnotes. Avoid explanatory footnotes by incorporating their content in the text. However, for essential footnotes, identify them with consecutive superscripts, as *book,*[2] *study,*[3] etc., and list the footnotes in a section, entitled FOOTNOTES, at the end of the text, but preceding the REFERENCES.

References. References should be listed alphabetically according to the author's last name at the end of the manuscript. Each entry should be numbered. Citation of a reference in the text should be made with this number, as (2); or if specific pages are to be indicated, as (2:245–7).

Illustrations of article and book references:

1. Bittner, Reign H. and Wilder, Carlton E., "Expectancy Tables: A Method of Interpreting Correlation Coefficients," *The Journal of Experimental Education,* 14:245–52, March 1946.
2. Thomson, Godfrey, H., *The Factorial Analysis of Human Ability,* Houghton Mifflin Co., Boston, 1950, 383 pp.

PROCEDURES

Send manuscripts to John Schmid, Department of Research and Statistical Methodology, University of Northern Colorado, Greeley, CO 80631.

Each contributor will receive 2 complimentary copies of the issue in which his article appears. Keep an exact carbon copy of your manuscript so that if a question arises the editors can refer you to specific pages, paragraphs, or lines for clarification by letter.

THE JOURNAL OF EXPERIMENTAL EDUCATION

4000 Albemarle Street, N.W., Suite 302,
Washington, D.C. 20016

Return Postage Guaranteed

Second Class
Postage Paid at
Washington, D.C.

THE *Journal* OF
Experimental Education

Volume 44, Number 2

Winter 1975

In this issue:

The Stability of Three Indices of Relative Variable Contribution in Discriminant Analysis

by Carl J. Huberty

The Educational Forces Inventory: A New Technique for Measuring Influences on the Classroom

by Nicholas F. Rayder and Bart Bödy

Teacher Classroom Management Skills and Pupil Behavior

by Walter R. Borg, Philip Langer, and Jeanette Wilson

THE JOURNAL OF EXPERIMENTAL EDUCATION

THE *Journal* OF EXPERIMENTAL EDUCATION

Volume 44, Number 2 Winter 1975

CONTENTS

The Journal of Experimental Education is published four times a year by HELDREF publications, 4000 Albemarle St., N.W., Washington, D.C. 20016. Annual subscription rates are $12.50 for institutions and $10 for individuals, plus $3 postage for all subscriptions outside the United States and Canada. Single copies $3. Second class postage paid at Washington, D.C.

All business correspondence should be sent to this address. Claims concerning missing issues made within 6 months will be serviced free of charge. Send all manuscripts to Prof. John Schmid, Department of Research and Statistical Methodology, University of Northern Colorado, Greeley, CO 80631.

Arvil S. Barr, Founder

EDITOR AND PUBLISHER • 1932-1962

(The Journal of Experimental Education is indexed/abstracted in Abstr. S. W., CSPA, Current Contents, Ed. Adm. Abst., Educ. Ind., Soc. of Ed. Abst., Current Index to Journals in Education, Language and Language Behavior Abst.)

VERBAL LEARNING AND SELF- AND SUPER-IMPOSED ORGANIZATION

J. WILLIAM MOORE
WILLIAM E. HAUCK
Bucknell University

JOHN FURMAN
Florida State University

ABSTRACT

The study was designed to test the general hypothesis that the acquisition of information is greater when the learner is trained to use a self-imposed organizational system rather than one which has been super-imposed on him by others. The results supported the hypothesis with a significant main effect for type of organization in favor of self-imposed training. In addition, it was found that with regard to the acquisition of information, serial memorization is more effective than self-imposed or super-imposed organizational systems in which no training is given; self-imposed testing situations are superior for retention; and self-imposed learning tested in self-imposed testing situations is superior to all other combinations of organization and testing.

THE PRIMARY PURPOSE of this research was to investigate the short- and long-term retention effects resulting from training students to organize and to classify information to be acquired.

A number of investigators have demonstrated that the retrieval of acquired information is enhanced by the organizational system used in the process of acquisition (1:331-335; 3:40-48; 1). It has been demonstrated also that self-imposed organizational systems, i. e., ones developed by learners themselves and used in the process of acquiring information, are more productive than organizational systems super-imposed on the learner (2:126-131).

In the present study, it was assumed that with regard to the acquisition and retention of information: (1) the use of self-imposed organizational systems is superior to super-imposed systems; and (2) individuals can learn to improve the quality of an organization which they create. Based on these assumptions, the general hypothesis was tested that training students to organize information which they are to learn increases their learning productivity and efficiency.

Method

Subjects

One hundred twenty eighth and ninth grade public school children of low-income families from a small mining community in central Pennsylvania participated in the experiment. The IQ range of the *Ss* was from 90 to 106, with a mean IQ of 100 and a standard deviation of 3.14.

Materials

The task which produced the data for the experiment and which was required of all *Ss* was to learn a list of 25 words predetermined by *E*. The only materials necessary were 3 x 5 cards with a single word printed on each card. The words were common nouns or verbs, i. e., block, nature, and sign, well known to the *Ss* and capable of being classified into several different organizational systems. For example, "block" could be placed in the categories of wood product, plaything, obstruction; "sign" in categories such as writing behavior or symbols, etc. All 25 words, then, were such that if *Ss* were asked to place them into seven or less categories, for example, several different classifications of categories were possible. Six words, in addition to the 25 used for the various treatments, were used for a practice trial.

Design

The design included five treatment groups labeled: self-imposed organization (SI); super-imposed organization (SUP); serial memorization (SM); controlled self-imposed organization (CSI); and controlled super-imposed organization (CSUP). All *Ss* were given the task of learning the list of 25 words predetermined by *E* discussed under *Materials*. The *Ss* in the SI group received training in the use of their own self-imposed word organization. Those in the SUP group received training by using a set of *E*'s word categories for organization. The SM *Ss* received training in serial memorization. The amount of training necessary for each *S* to reach a criterion of consistent replication of a single organization of words, whether it be self- or super-imposed, varied. To control for the varying amount of training across *Ss*, the CSI group experienced the *average* amount of the same training as the SI group; and the CSUP group received the *average* amount of the same training as the SUP group.

After training *Ss* with either a self- or super-imposed organization, it was possible to test for the retention of the 25 learned words by asking *Ss* to classify as many of them as they could recall into either the organizational scheme on which they were trained or a different one. Thus it was possible to test for retention by using either *E*'s or *S*'s organizational systems. A testing dimension of two levels, self-imposed (SIT) and super-imposed (SUPT), was included in the design. The effect of retention was examined by including two levels in the design, short- and long-term retention.

The *S*'s were assigned to all treatments in a 2 x 2 x 5 factorial with: two levels of retention, short and long; two levels of type of test, self-imposed (SIT) and super-imposed (SUPT); and five levels of treatment, SI, SUP, SM, CSI, and CSUP. Originally, the experiment was carried out with 60 *Ss*; later, 60 additional *Ss* drawn from the same population were made available and the experiment was replicated. Thus, the final analysis was completed using a replicated design of two 2 x 2 x 5 factorials (6:391-394).

The retention test required *Ss* to recall as many of the 25 words as possible. The measure of short-term retention was administered immediately following the termination of the experimental procedure, and the measure of long-term retention was administered after an interval of 48 hours following the termination of the experimental procedure.

Procedure

In one practice trial, the *Ss* in the five experimental groups experienced an example of what was required by using the practice cards in a way appropriate for their experimental condition. Following the example, questions of the *Ss* were answered with respect to the procedure followed. Then the experimental condition appropriate for the respective groups was implemented by using the 25 remaining word-cards.

The SI group was instructed to put the words appearing on the cards into columns any way they wished by using no less than two and no more than seven columns. They were also instructed to give a reason for the organization they used on each trial. The words were rearranged randomly after each trial and were presented again for the *Ss* to organize. These *Ss* continued the experimental task until they attained two successive identical sorts.

The SUP *Ss* were given by *E* five headings naming categories into which the words were to be placed. On the first trial, *E* showed the *Ss* which words were to be placed under the respective category headings. Each word was randomly chosen and placed in its category, and a rationale was then given by *E* for that particular placement. The *Ss* were then presented with the cards in random order and told to put them into the proper categories. All mistakes were corrected at the end of the trial. *Ss* had the task of placing all the words under the correct headings until two successive, successful trials were achieved.

The SM group acted as a control for the training in organization. Each *S* in this group was presented with the words in serial fashion for four trials, the average number of trials required for both the SI and the SUP groups together.

The average number of exposures to criterion required for the SI group was three; therefore, each *S* in the CSI was exposed to the SI procedure three times. The *Ss* in the CSI group were asked to categorize the words any way they wished by using two to seven columns.

The average number of exposures to criterion required for the SUP group was five; therefore, each *S* in the CSUP received five exposures to the words. The *Ss* in the CSUP group were asked to categorize the words by using *E*'s organizational scheme.

Results and Discussion

An ANOVA with method, type of test, type of retention, and replication was completed with the number of words correctly recalled as the dependent variable (see Table 1).

The main effect of method was significant ($p < .01$). A Newman-Keuls post-test revealed that the groups using the self-imposed method of organization (SI), super-imposed organization (SUP), and serial memorization (SM) differed significantly from both the self-imposed control group (CSI) and the super-imposed control group (CSUP), ($p < .01$). Inspection of the means (see Table 2) indicates that the means of the CSI and the CSUP groups were lower than the other groups.

What is most interesting about these findings is that the SM group performed higher than either the CSI or CSUP group, suggesting that the SI and SUP organization without training may enhance recall less than SM, or that training may contribute more to recall than the organizational style alone.

As can be observed in Table 1, the main effect of tests was also significant ($p < .01$). The *Ss* using the SIT testing condition recalled more words (17.20) than the *Ss* using the SUPT testing condition (16.65). This observation suggests that a super-imposed organizational system applied to testing differs from a self-imposed organizational system in that a super-imposed testing situation may interfere with the recall of information acquired and thus produce a lower performance as happened here.

The effects of the testing condition on long-term retention are apparent in the significant interaction ($p < .05$) observed between testing conditions and short- and long-term retention. A Newman-Keuls post-test indicated a significant difference ($p < .01$) between the SIT and the SUPT condition for long-term but not short-term retention. The mean of the SIT condition (15.93) was greater than the mean of the SUPT condition (13.33). This observation supports the notions that the testing condition serves as a learning situation and that the presence of a SI condition in testing where long-term retention is a consideration is critical to performance.

Table 1.—ANOVA of the Number of Correct Words Recalled

Source	df	MS	F
Treatment (Tr)	4	85.32	11.29**
Test (T)	1	72.07	9.53**
Retention (R)	1	385.20	50.95**
Replication (Rep)	1	5.20	<1.00
Tr x T	4	27.60	3.65**
Tr x R	4	12.15	1.61
Tr x Rep.	4	14.15	1.87
T x R	1	33.08	4.38*
T x Rep.	1	46.88	6.20
R x Rep.	1	15.42	2.04
Tr x T x R	4	11.43	1.51
Tr x T x Rep.	4	5.56	<1.00
Tr x R x Rep.	4	18.72	2.48
T x R x Rep.	1	14.00	1.85
Tr x T x R x Rep.	4	7.74	1.02
Error	80	7.56	

**p<.01

*p<.05

Table 2.—Means and Standard Deviations for Treatment

Treatment	Mean	SD
SI	18.33	2.13
SUP	17.46	3.23
SM	17.46	4.01
CSI	14.96	3.62
CSUP	13.92	4.11

Table 3.—Means and Standard Deviations of the Treatment by Test Interaction

	Treatment									
	SI		SUP		SM		CSI		CSUP	
Test	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
SIT	20.66	2.35	17.17	3.22	18.08	4.22	16.25	3.77	13.83	4.11
SUPT	16.00	1.93	17.75	3.38	16.83	3.98	13.67	3.63	14.00	4.29

The ANOVA also indicates that the main effect of retention was significant ($p < .01$), with the *Ss* in the short-term recall groups recalling a greater number (18.22) than the *Ss* in the long-term recall groups (14.65). This observation is consistent with the literature concerning short- and long-term recall.

Probably the most interesting finding was the significant interaction ($p < .01$) of treatment and type of test. The Newman-Keuls post-test revealed that SI *Ss* differed significantly ($p < .05$) from all other groups under the student-initiated testing condition (see Table 3). This observation is consistent with the findings of Tulving (5:219-237) and Mandler and Pearlstone (2:126-131) who have demonstrated that self-imposed organizational systems enhance the recall of information acquired. Further, it extends the research in this area by demonstrating that training *Ss* to organize information with the development of their own organizational system as a goal provides a more effective retrieval system than either imposing organizational systems or simply encouraging but not training *Ss* to develop their own organizational system.

The fact that the SI group differed from the SUP and the SM groups only under the self-imposed testing conditions supports the earlier suggestion that the super-imposed testing condition may interfere with the existing organizational system of the learner and thus reduce its effectiveness as a retrieval system of acquired information. This observation is similar to those made by Mandler and Pearlstone (2:126-131) when they forced their *Ss* to learn other *Ss*' methods or organizing word lists.

It was also of interest to note that the mean of the SUP group was significantly different from that of the CSI group under the super-imposed testing conditions in that the mean of the SUP was the greater; however, no significant difference occurred between the two for the self-imposed testing condition. A probable explanation for this observation is that the super-imposed testing condition interferes with recall to a greater extent under SI conditions than under SUP conditions. An inspection of the means (see Table 3) supports this explanation.

The importance of a self-imposed organizational system as a result of training is also emphasized by the fact that the SM group differed ($p < .05$) under the self-imposed testing conditions only from the SI group and the CSUP group, with the mean of the SM group being less than the mean of the SI group and greater than the mean of the CSUP group.

In summary, the results of this experiment tend to support the hypotheses that student-initiated organizational systems, both in the organizational and testing situations, contribute to the recall of information, and that the training of students to develop their own organizational systems increases their effectiveness in the long-term recall of acquired information. Assuming the validity of these findings, it is recommended that classroom teachers provide greater emphasis on providing conditions necessary for their students to acquire a concept of organization of information especially where long-term recall is the consideration.

REFERENCES

1. Ausubel, David P., "A Cognitive Theory of School Learning," *Psychology in the Schools*, 6:331-335, October 1969.
2. Mandler, George; and Pearlstone, Zena, "Free and Constrained Concept Learning and Subsequent Recall," *Journal of Verbal Learning and Verbal Behavior*, 6:126-131, April 1966.
3. Postman, Leo J., "Does Interference Theory Predict Too Much Forgetting?" *Journal of Verbal Learning and Verbal Behavior*, 2:40-48, July 1963.
4. Tulving, Endel, "Subjective Organization in Free Recall of 'Unrelated' Words," *Psychological Review*, 69:334-354, July 1962.
5. Tulving, Endel, "Intratrial and Intertrial Retention: Notes Toward a Theory of Free Recall Verbal Learning," *Psychological Review*, 71:219-237, May 1964.
6. Winer, Ben J., *Statistical Principles in Experimental Design*, McGraw-Hill, New York, 1971, 907 pp.

INTERPRETING CONSERVATION TRAINING: A DEVELOPMENT ON HOFMANN

D. J. F. MÜLLER
Lancaster University, England

ABSTRACT

A refinement of Hofmann's paradigm for conservation training is proposed. It is suggested that a delayed learning task be utilized for assessing generalization. This, it is argued, would also be a more preferable measure of whether or not new cognitive structures have been acquired.

THE MOST DIFFICULT PROBLEM in assessing conservation training is satisfying the criteria laid down by Piaget (2) to demonstrate that development rather than simple learning has occurred. As Hofmann (1) notes, this involves ascertaining that the change is stable over time, that it is dynamic in that it can lead to generalizations, and that more complex structures have been acquired.

From an empirical point of view, the most difficult of these criteria to operationalize is that of testing for the acquisition of more complex structures. Within Hofmann's paradigm, the original pre-test is used to assess whether or not a cognitive structural change has occurred. However, as Hofmann clearly shows, it may be misleading to credit any improvement on this test to a change in the child's cognitive structures, since such improvement may simply be the result of an interaction between the pre-test and the training which has produced only rote learning.

As a control for this, Hofmann incorporates a delayed post-test into his design to assess whether or not any improvement on the immediate post-test is merely the result of external reinforcement.

Criterial Problems

Two criteria must be satisfied for it to be suggested that more complex structures have been acquired. Hofmann suggests that: (1) there should be a difference between the delayed post-tests of the trained and nontrained groups; and (2) there should be no difference between the immediate and delayed post-tests of the trained group. Failure to meet these criteria, even though there may be differences on the immediate post-tests, is interpreted as a possible result of rote learning.

These criteria, though valid, are not essential for suggesting that more complex structures have been acquired. It is not critical that the trained group's performance be the same on the immediate and delayed post-tests and better on the delayed post-test than the control group. Any decrement shown by the trained group from the immediate to the delayed post-test, and any failure to maintain a higher score than the control group might simply be the result of extraneous variables.

One crucial uncontrolled variable might be memory. It is not possible to deduce on the basis of a decrement in performance on an immediate or on a delayed post-test that there is therefore an *a priori* lack of cognitive structures. The crucial question to be considered is whether or not material with the same underlying structure can be assimilated, that is, incorporated into any such structures if they exist, and *not* whether or not such material can be remembered. One is concerned with a qualitative difference in learning, not a quantitative difference in memory. On the basis of this distinction, one can suggest that if new structures have been formed, the trained group should be better equipped to learn and assimilate such material.

A Possible Solution

A more preferable measure of whether or not any more complex structures have been acquired might be a learning task involving the same logical structures as those trying to be induced in training. Performance on this learning task would effectively assess whether or not the training had produced any structural changes. However, one cannot use a learning task as a repeated-measures design for assessing pre- and post-test performance. It is necessary, then, to design a learning task independent of the post-test but related to the underlying structure of the training schedule.

One solution to this problem is designing the generalization task such that it can be utilized as a learning experience. In many situations this would be a more preferable test of whether or not cognitive structures have been acquired. First, it recognizes that the timelag on immediate and delayed post-tests allows other variables to interfere

in the process. Second, it stresses the qualitative aspects of cognitive growth by assessing not simply whether material can be better remembered but, rather, whether new material can be learned or assimilated into the subject's cognitive network. Furthermore, by delaying the giving of this generalization task until after the required time delay, one can assess independently of the delayed post-test whether the learning is stable over time. This design may be represented schematically as suggested by Hofmann.[1]

A carefully designed generalization task can provide data relevant to all three questions posed by Piaget. It provides a measure of whether the learning can be generalized, and whether it is stable over time, and also enables a qualitative assessment of whether or not new cognitive structures have been acquired.

Hofmann stated that he made no attempt to cover all the possible error routes or outcomes associated with conservation training. He did, however, outline a viable experimental framework capable of being developed. A development suggested in this paper is that for many cases it may be preferable and sometimes essential to make better and more extensive use of the generalization task in assessing training procedures. This is the case particularly when the critical variable is the acquisition of cognitive structures. Hofmann's paradigm has been developed to give more detailed consideration to the most difficult Piagetian criteria to operationalize.

FOOTNOTE

1. Utilizing the same formula and notations as Hofmann (1:49) gives the following research design:

$T[1]$	X	$T[2]$		$T[3]$	$T[3]A$
$T[1]$	X	$T[2]$	(at least	$T[3]A'$	$T[3]'$
$N[1]$		$N[2]$	two weeks)	$N[3]$	$N[3]A$
$N[1]$		$N[2]$		$N[3]A'$	$N[3]'$

Thus, in the training conditions, the pre-test is followed by training and the post-test. After two weeks the delayed post-test is given first and then followed by the generalization task, or vice-versa. In a similar way, two control groups are formed.

REFERENCES

1. Hofmann, R. J., "A Piagetian Paradigm for Interpreting and Generating Research Dealing with Conservation Training," *Journal of Experimental Education,* 43:47-52, Fall 1974.
2. Piaget, J., "Development and Learning," in R. E. Ripple; and V. N. Rockcastle (eds.), *Piaget Rediscovered: A Report on the Conference on Cognitive Studies and Curriculum Development,* Cornell University Press, Ithaca, March 1964.

THE EFFECTS OF CONTINUOUS PROGRESS INSTRUCTION IN A COLLEGE RELIGION COURSE

J. WILLIAM MOORE
Bucknell University

ELLEN D. GAGNE
University of Wisconsin

WILLIAM E. HAUCK
Bucknell University

ABSTRACT

While a previous study has shown that requiring a criterion of mastery in a hierarchically related discipline, i. e., Physics, leads to improved transfer (as well as improved acquisition and retention), it is not known whether a greater knowledge of prerequisites in a less hierarchically related discipline, i. e., Religion, leads to improved transfer. The findings reported here which show no differences for transfer as a function of instructional procedure in Religion lead to the conclusion that the learning of prerequisites may have more significant effects on transfer only in the more hierarchically related disciplines. The findings for requiring mastery in a Religion course indicate that acquisition is facilitated in this discipline when the objectives of the course include either analysis and interpretation or factual recall.

WHILE INSTRUCTIONAL DEVELOPMENT and evaluation has proliferated at the pre-college level (e. g., SMSG Math, SRA Science, Criterion Reading), less has been done to develop and evaluate instructional systems at the college level. One attempt that has been made at the college level, developed as part of a larger total system of institutional improvement (4), is called the "Continuous Progress Plan." The Continuous Progress concept is derived from two basic psychological considerations: (a) that there are individual differences in learning rate, and (b) that an increase in the probability of an undesirable response occurs when the receipt of a desired reinforcer is made contingent on the occurrence of an undesirable response. In Continuous Progress courses provision is made for individual differences

in learning rate by allowing students to decide, within certain limits, when they will be evaluated. In addition, desirable progress and ultimate success in a course are ensured by making them contingent on reaching criterion levels of mastery in succeeding units of material.

Previous controlled experiments conducted to evaluate the Continuous Progress concept of instruction have found greater acquisition and more positive attitudes for Continuous Progress students than for traditionally taught students in Psychology, Philosophy, and Biology courses (3) and in a Physics course (5). One purpose of the present study was to extend these findings to another discipline, Religion. It has been shown for Physics that there is greater transfer to a later Physics course and greater retention one year after original learning as a function of Continuous Progress instruction taught students in Physics (5). A question which is raised by the finding of greater transfer for Continuous Progress students is whether transfer for Continuous Progress would occur in a discipline less hierarchically related than Physics, i. e., Religion. While it would be expected from a theory of hierarchically related knowledge structures (2) that greater original learning in prerequisite courses facilitates transfer in later related courses (e. g., Physics), this should not be the case in less hierarchically related disciplines such as Religion.

A final question posed by this study involved the types of educational objectives that can be more efficiently attained using Continuous Progress procedures. In previous studies, objective questions emphasizing either recall (Biology, Psychology, Philosophy) or problem solving (Physics) were employed. Can students also achieve analysis and interpretation objectives (measured through essay items) more easily in a Continuous Progress course? The second year of the Religion experiment reported here addressed itself to this question.

Method

Subjects

A total of 46 students was evaluated—30 in 1968-69 and 16 in 1969-70. Students were randomly assigned to Continuous Progress (CP) or Control (C) sections of the course. Mean verbal SAT scores were not significantly different from CP and C groups in 1968-69 (t = .14) or 1969-70 (t = .31).

Teaching Method

Both C and CP sections were taught by one instructor in 1968-69, and another instructor taught both C and CP sections in 1969-70. Course textbooks and objectives were the same for both sections of the course each year, although the objectives for 1968-69 emphasized factual recall, while the objectives for 1969-70 included both recall and high-order objectives such as comparison and evaluation of different writers' viewpoints. For example, test items for the 1968-69 course were of the following general nature:

Cicero applied the term "religo" to:

a. national customs
b. a celestial power
c. family rites
d. an inner attitude
e. probable superstitions

In 1969-70, however, test items included some comparative analysis, as examplified in the second half of this question: "What does Smith mean by 'cumulative tradition'? How is it related to faith?"

Control *Ss* both years were instructed to read the appropriate text materials and attend three lecture-discussion sections each week. All C subjects took unit exams at the same predetermined time. CP subjects, while receiving the same assignments as C subjects, did not attend lectures, but instead were informed of the objectives of a unit of material and instructed to come to the professor to take the exam when they thought they were ready. (The likelihood of "cheating" was minimized by having several alternate forms of the test available for individual students.) If a CP subject did not reach the satisfactory criterion (80% correct) on his first attempt at a unit test, he reviewed his errors with the instructor, studied some more, and took an alternate form of the unit test when ready. This test-review-retest cycle was repeated, when necessary, a number of times until *Ss* attained the 80% criterion.

Evaluation

In 1968-69 both C and CP groups took the same ten unit tests and the same final. In 1969-70 both groups took six common unit tests and a common final. Comparisons of C and CP groups' scores on the unit tests and finals were used for the evaluation of acquisition.

The grades received in the first Religion course taken after completion of the CP Religion course (or its corresponding traditional form in the case of the C groups) were used to evaluate transfer effects both years. A questionnaire assessing student perception of the procedure was administered to the 1968-69 group in which the following questions were posed:

1. Do you feel that a minimum level of achievement as a requirement for proceeding to the next unit is a desirable requirement?
2. Do you believe that the procedure for permitting a student to proceed at his own rate is desirable?

Results and Discussion

Acquisition

The median score and interquartile range for unit tests (summated over tests) and final exams each year are shown in Table 1. For 1968-69, both first and last attempts to reach a criterion score on unit tests are included for the CP group, while for 1969-70, the first attempt was not available for analysis. Also for 1969-70, the final two (of six) unit tests were not included in the summated score

Table 1.—Central Tendency and Variance of Unit Test and Final Exam Scores

Score	Group	Median	Interquartile Range
		1968-69	
First Attempt (summed over units)	CP	500	26.00
	C	324	72.00
Last Attempt (summed over units)	CP	511	20.00
Final Exam	CP	70	6.25
	C	52	9.50
		1969-70	
Last Attempt (summed over units)	CP	362	10.75
	C	317	67.63
Final Exam	CP	75	8.13
	C	70	13.38

for unit tests because two *Ss* in the C group dropped the course after the fourth unit test, thus introducing bias.

For 1968-69, Mann-Whitney *U* tests revealed significant differences between the CP group's first attempt to reach criterion on the unit tests and the C group's unit tests ($z = 4.76$; $df = 13, 23$; $p < .001$), the CP group's last attempt and the C group's unit tests ($z = 4.92$; $df = 13, 23$; $p < .001$), and the CP and C groups' final exam scores ($z = 3.63$; $df = 7, 23$; $p < .01$). In each of these comparisons, the CP group showed the higher performance.

The finding of higher performance even on the first test is like the finding of higher performance on the first test for CP Biology, but unlike the finding on the first test for Psychology (3) of no significant difference between C and CP groups. In two of three courses, the intuitive expectation that CP students would not prepare for their first attempt, but rather take the first test to find out the nature of the test, is clearly unfounded.

For 1969-70, Mann-Whitney *U* tests revealed significant differences between CP unit tests (last attempt) and C unit tests ($U = 2$; $df = 7, 9$; $p < .002$), but not between CP and C groups' final exam scores. For unit tests, the CP group showed higher performance than the C group.

Two explanations for the finding of no differences on the final exam seem plausible. First, two C subjects who obtained low grades on the first four tests dropped the course, while no CP students dropped. The result could be an inflation of C group scores relative to what they would have been had the two poorer students remained in the course. Secondly, it is possible that CP subjects did not study hard for the final because of the assumption (correct for most of them) that since they had above a B average going into the final, they could afford to do less than B level on the final and still get a B in the course. If this is the case, a different set of contingencies (e. g., weighting the final more heavily) could be constructed to maintain CP performance at a higher level even on the final.

It is interesting to observe that the semi-interquartile range of the CP group (8.13) is considerably lower than that of the C group (13.38). This suggests that the treatment is effective in reducing variability.

Since the 1969-70 course used essay exams, the results seem to warrant the conclusion that Continuous Progress procedures facilitate acquisition of higher-order objectives such as analysis and comparison (1969-70) as well as recall (1968-69) objectives. This conclusion is limited, however, by the fact that inspection of the essay questions used in 1969-70 revealed more recall questions than comparison- or analysis-type questions. Further research on the question of types of objectives which may be attained efficiently using CP procedures needs to be conducted where test items are carefully written to measure various types of learning.

Transfer

A Mann-Whitney *U* test completed on grades received in the first Religion course taken following completion of the CP and C sections of the experimental course revealed no significant difference between C and CP groups ($U = 90$, ns). Thus, it appears that while CP procedures facilitate acquisition in a Religion course, they do not improve transfer.

Since a previous study (5) found greater transfer in Physics for a CP group than for a C group, it is plausible

that CP procedures facilitate transfer only in courses in which a strong hierarchical relation exists between the CP course and the more advanced course. That is, greater acquisition shown by CP subjects will facilitate acquisition at a more advanced point in the hierarchy.

Students' Perceptions

Mann-Whitney *U* tests (normal approximation) of the two questions regarding perceptions of CP procedures revealed no significant differences between groups. Both groups were neutral ($z = 1.27$, ns) toward the requirement of mastery, and positive ($z = 1.57$, ns) toward the procedure of allowing students to progress at their own rate. While the direction of perceptions for both groups is not surprising, it is interesting that the CP group was not significantly more positive toward CP procedures than the C group, as had been found previously (3). A unique feature of the 1968-69 Religion course is that the instructor became convinced early in the semester that CP procedures would not facilitate learning and subsequently communicated this to the CP subjects. Despite greater success, perhaps their perception was not more positive than the C group's because of the negative attitude toward CP procedures held by the instructor.

Summary and Conclusions

In summary, the findings of this study add to the generality of the concept of Continuous Progress instruction by showing that (a) acquisition is improved in Religion, and (b) acquisition is improved for higher-order objectives. The finding of no transfer effects for Religion tends to support the argument that CP procedures facilitate transfer only in courses where long hierarchical relations exist.

REFERENCES

1. "A Continuous Progress Plan: Bucknell's Experiment Aims for Mastery," in *College Management*, January 1968, pp. 12-20.
2. Gagne, R. M., *The Conditions of Learning*, Holt, Rinehart and Winston, New York, 1965.
3. Moore, J. W.; Mahan, J. M.; and Ritts, C. A., "Continuous Progress Concept of Instruction with University Students," *Psychological Reports*, 25: 887-892, 1969.
4. Moore, J. W., "An Evaluation of the Cost and Quality of Instruction: A Plan for Instructional Improvement," unpublished paper presented at the University of Brazil, Rio de Janeiro, 1971.
5. Moore, J. W.; Hauck, W. E.; and Gagne, E. D., "Acquisition, Retention and Transfer in an Individualized College Physics Course," *Journal of Educational Psychology*, 64: 335-340, June 1973.

EFFECT OF TRAINING IN THE USE OF BEHAVIORAL OBJECTIVES ON STUDENT ACHIEVEMENT[1]

RONALD E. BASSETT
The University of Texas at Austin

ROBERT J. KIBLER
The Florida State University

ABSTRACT

This study was an experimental investigation of the effects on cognitive learning of training students to use behavioral objectives. One half of the *Ss* received training via programmed instruction in the use of objectives. They were also required to achieve a criterion score on a measure of ability to use objectives. The remaining *Ss* received a placebo treatment. Results indicated that *Ss* receiving training achieved statistically significant higher scores on an examination consisting of items matched to objectives than *Ss* not trained, although the absolute difference gave no support to useful practical application.

SUPPLYING EXPLICIT STATEMENTS of instructional objectives to learners is an integral aspect of mastery learning models of instruction (1, 2, 5). This practice appears to be based on the assumption that objectives will reduce the student's uncertainty about what is required of him, thus permitting the student to maximize learning by selectively attending to the most relevant stimuli in the instructional setting. If this assumption is valid, then it is reasonable to expect that when performance is compared between *Ss* given behavioral objectives (BOs) and *Ss* not given objectives, those possessing objectives should exhibit greater learning. As Duchastel and Merrill (6) demonstrated in their extensive review of objectives research, however, this relationship has not been *consistently* observed. While

learner posession of objectives has been shown to facilitate learning in a number of studies, such a facilitating effect has not been observed across *all* studies. The generalizability of such an effect is therefore quite difficult to determine at this time. Furthermore, serious methodological problems appear in the literature with such frequency that it is possible to place reasonable confidence in few of the studies.

Although there are many methodological inadequacies in the objectives literature, the ability of learners to use the objectives given to them emerges as an especially critical question for research. Several investigators have suggested that students need to know how to use objectives before effects on learning will be present (3, 4, 9, 12, 13). Only three studies have been found, however, which report procedures for training learners to use objectives. Boardman (3) and Brown (4) attempted to train *Ss* to use BOs, although neither assessed the effectiveness of the training. Furthermore, both concluded from anecdotal evidence that their limited training procedures were probably inadequate. In contrast, Morse and Tillman (12) empirically tested the effects of their training efforts.

Morse and Tillman's training consisted of having students read Mager's (11) *Preparing Instructional Objectives* with accompanying classroom instruction. *Ss* in a second training condition were directed to read Mager's book out of class, with no classroom instruction provided. A third group (control) was directed to perform an unrelated task.

In the second part of the study, one half of the *Ss* were given BOs for an assigned reading and the remaining half were not given objectives. *Ss* with objectives achieved higher scores on test items matched to those objectives than did *Ss* not posessing objectives. However, no significant main effect due to training and no significant interaction effect between training and possession of BOs were found. Consequently, Morse and Tillman concluded that training was not necessary for students to use objectives effectively in learning.

The factor which most seriously limits the confidence which may be placed in this conclusion concerns the validity of the training procedures. Morse and Tillman acknowledge that Mager's book provides information about objectives, but does not contain instruction in how to *use* objectives in learning. Hence, the validity of the training is questionable. Conclusions about the effects of training cannot be drawn without establishing a strong correspondence between the training and the required terminal behavior.

Since BOs are assumed to be a learning tool, it seems reasonable that students may require training before they are able to use objectives with maximum effectiveness. However, most investigators have ignored the question of student ability to use objectives on the assumption that when learners are given BOs they will use them, and that they will use them as the investigator intended. Because little information on the need for training students to use objectives is presently available, the validity of the two assumptions is not known. However, if training is necessary, and it is not provided, then positive effects of BOs may not emerge. Because it seems important to determine if training learners to use BOs is necessary, the relationship between training in the use of objectives and learner achievement was investigated in this study. Specifically, this hypothesis was tested: When objectives are provided for a unit of instruction, *Ss* trained to use objectives will achieve a significantly greater number of correct answers on an examination consisting of items matched to the objectives than *Ss* not so trained.

Method

Subjects

Ss ($N = 159$) were undergraduate students enrolled in a survey course of human communication theory at a major southern university. *Ss* varied extensively in the fields selected for their major(s) and minor(s). They were not informed that a study was being conducted.

Training

During a previous term students enrolled in the course who achieved at least the grade of 'C' completed a questionnaire which asked them to identify the steps they went through in using objectives to study for course examinations. From these self-reports, five steps in using objectives were identified:

1. Read the objective to identify where important material may be found.
2. Read the material to locate specific passages related to the objective.
3. Read the objective to determine the form of the test item.
4. Read the objective to determine what you must be able to do to answer the test item correctly.
5. Ask yourself a question in a form similar to the one you will be asked on the test and try to answer it.

From these five steps, the following six objectives for training students were derived:

1. Given a behavioral objective in which one part(s) of the objective (e. g., the part identifying where important material can be found) is underlined, and five alternative statements, the student will select the statement which most accurately describes why the underlined part(s) of the objective is valuable when using BOs to learn.
2. Given a behavioral objective, a reading passage divided into five separate, numbered parts, and five alternatives from which to choose, the student will select the alternative which correctly identifies the part(s) of the reading passage which contains information most relevant to the behavioral objective.
3. Given a behavioral objective, a reading passage which contains information relevant to the objective, and a sample multiple-choice test item matched to the BO, the stu-

dent will select from the five alternatives in the sample test item the correct answer to the item.

4. Given a behavioral objective which has been divided into numbered parts, and five alternatives from which to choose, the student will select the alternative which correctly identifies the part(s) of the BO specifying: (a) the form of the test item; (b) what the student must do to answer the test item correctly; and/or (c) where important material can be found.

5. Given the five steps in analyzing/using behavioral objectives, and five alternatives from which to choose, the student will select the alternative which correctly: (a) identifies the five steps in proper sequence from first to last, or (b) identifies the proper place of any step(s) in the sequence.

6. Given a behavioral objective and three sample test items, the student will select the best item(s) which is most closely matched (i. e., most appropriate) to the test item form specified by the behavioral objective.

A 60-frame, branching type, instructional program was developed to teach *Ss* how to perform each behavior. Examples of BOs and test items appearing in the program (as well as the training tests) were drawn from the various units which composed the course (i. e., intrapersonal; interpersonal; small groups; nonverbal; and mass media communication). The program underwent three separate revisions on the basis of responses obtained in a pilot study. The validity of the training is supported because the behaviors which the program was designed to teach were derived from strategies successful students reported employing in using BOs to learn.

Training Tests

Four test forms were developed. Each test form contained at least two test items for each of the six objectives for the programmed instruction. The first form contained 20 items and the remaining three forms each consisted of 12 items. The minimum level for acceptable performance was set at 90% correct answers for each test form.

The validity of the training tests was assessed by having six trained judges rate on a three-point scale the extent to which each of 21 items randomly selected from the four test forms corresponded to the objectives to which they were matched. Perfect correspondence between the 21 items and the matched objectives would be represented by a summated score of 63. The mean of the summated scores for the six judges was 61.33. This value indicates high direct validity of the items. The inter-rater reliability of these ratings obtained by Ebel's (7) analysis of variance procedure was .98.

Reliability coefficients obtained by use of the Kuder-Richardson 20 for the four forms were .71, .63, .76 and .03.[2] Coefficients obtained by Livingston's (10) criterion-referenced procedure were .92, .69, .77 and .17.

Unit Examination

The dependent variable in the investigation was the number of correct answers obtained by *Ss* on a 28-item multiple-choice (five alternative) test. The test consisted of two items for each of 14 objectives constructed for Kenneth Gergen's (8) *Concept of Self.* Objectives were written in the format of this sample:

> Given five alternative statements, the student will select the statement which most accurately illustrates or describes the concept of *double bind* (Chapter III).

The following test item was written to match the objective:

> Select the alternative which best illustrates the concept of *double bind:*
>
> A. Martha and Milton decide they want to eat dinner out Friday night. She wants Greek food and he wants Hungarian food. They cannot agree on where to go.
> B. Armando's doctor tells him he has an ingrown nail that must be corrected now. Armando decides to wait until he can afford the expense.
> C. After having her color television repaired, Debbie pays the serviceman but is not satisfied with the way the machine works.
> D. Gina needs nine hours to graduate. She cannot decide whether to take one 5- and one 4- hour course, or three 3-hour courses.
> E. Harold's wife Louise tells him often that she loves him, but frequently ruins his favorite meals by overcooking them.

Six trained judges examined the test items and agreed unanimously that each item satisfied the specifications of the objective to which it was matched.

The reliability of the scores obtained on the test was determined to be .86 using the norm-referenced Kuder-Richardson procedure, and .92 using Livingston's (10) criterion-referenced procedure.

Procedure

To clarify the description of the administration of the experimental treatments, the events which took place during each of the first four class sessions are discussed in the temporal sequence in which they occurred.

First Class Session

Ss were randomly assigned to either the training or no-training treatments. Within each treatment condition, each *S* was randomly assigned to one of three instructional sections. Each section was supervised by two instructional assistants (*IA*s). Three graduate students and nine undergraduates served as *IA*s. Each of the undergraduates had completed the course in the previous term with the grade of 'A'. *IA*s were randomly assigned to the six instructional sections.

Second Class Session

All *S*s reported to their assigned sections. *S*s in the sections designated to receive training were informed of this requirement by the *IA*s, who also discussed general classroom procedures. The *IA*s then distributed Form I of the training test. *S*s were told that if they answered 90% of the questions correctly, they would not have to work through the programmed instruction nor take any additional forms of the training test. *S*s used machine scorable IBM answer sheets to record their answers. Upon completing the test, *S*s returned their test copies and answer sheets to their *IA*s. *S*s were told that attendance was required at the next class meeting, at which time they would be informed of their performance on the test.

Following class, the answer sheets were scored. Of the 76 *S*s completing the test, only three achieved the 90% criterion.

*IA*s in the three sections which did not receive training told *S*s about general classroom procedures and announced that attendance for the next class session was required.

Third Class Session

*S*s in the training sections were informed of their performance on Form I of the training test. *S*s who achieved the 90% criterion score were excused from class. *S*s not reaching criterion were given the programmed instruction, directed to work through it and to then request Form II of the training test.

When a *S* completed Form II of the training test, he returned the test copy and his answer sheet to his *IA*s, who immediately scored the answer sheet with a punched answer key and completed a feedback sheet which they gave to the *S*. The feedback sheet informed the *S* of the percentage of correct answers which he had obtained. If his performance was less than the 90% criterion level, the feedback sheet identified the BOs for the program which corresponded to the test items answered incorrectly. The feedback sheet also identified frames in the program which contained information relevant to the unmastered objectives. The *IA*s then returned the *S*'s copy of the program and encouraged him to restudy it. This set of procedures was repeated for *S*s who failed to achieve the criterion level for the third form of the training test. Only six *S*s failed to achieve criterion on the third form, but each of these was able to achieve the 90% level on the fourth form.

*S*s in the no-training condition reported to their assigned classrooms and received the following placebo treatment: They were informed by their *IA*s that a graduate student in Communication needed their assistance in conducting research. Each *S* was given a booklet which contained directions for completing semantic differential scales related to the nonverbal behavior of teachers in classroom settings. The task required approximately 45 minutes to complete. *S*s were informed that by completing the task they satisfied the course requirement that they participate in an experiment.

Before leaving class, *S*s in all sections were given copies of the BOs and required readings for the first unit. *S*s not given training were told they could take the test for the first unit at the next class session, if they wished to do so. *S*s receiving training were informed that they could take the examination for the first unit *only* after they had achieved criterion on the training test.

Fourth Class Session

*IA*s for all sections were present in their assigned classrooms to answer questions regarding the readings and objectives for the unit. *S*s were permitted to attempt the examination. Testing was self-paced, i. e., a *S* took the test when he felt sufficiently prepared. To attempt the examination, a *S* requested a test copy and answer sheet from his *IA*s. The *IA*s scored the answer sheet as soon as the test was completed. When scores were available for all *S*s on the examination, the data were analyzed.

Results

A directional *t*-test for independent data was computed for the number of correct answers achieved by the two groups on the course examination. The *t*-test analysis produced a significant *t*- value ($t = 2.37$; $df = 157$; $p < .01$), indicating that the trained *S*s had higher scores, thus supporting the hypothesis. Table 1 summarizes the results of the analysis.

Table 1.–Summary of Analysis

Summary data	Treatment	
	Training	No-Training
Mean	24.29	22.78
Standard deviation	3.33	4.51
Cell size	76	83

Discussion

Although differences in achievement between trained and untrained learners were statistically significant, the absolute difference between the mean scores of the two groups was less than two points. Such a small effect attributable to training might possibly be accounted for in at least two ways. First, *IA*s answered all questions which *S*s asked about the unit objectives and their relation to the unit test. Hence, although untrained *S*s did not receive formal instruction in the use of BOs, they were not denied information provided informally about their use. Second, although no data are

available, it seems probable that *Ss* receiving training discussed the instruction with some *Ss* not trained. Some untrained Ss may also have discussed the use of BOs with friends who were enrolled for the course in previous terms.

Since a significant postive effect for training was found despite conditions which might have mitigated the effect, it seems important that future investigations of learner posession of objectives should account for learner competence in their use. If it is indicated that learners lack the basic ability to use objectives, then training should be provided.

Additional research is needed to increase knowledge regarding: (1) what types of learning are facilitated by instruction in the use of objectives; (2) what the nature of such instruction should be, i. e., the most effective way to use objectives in various types of learning; and (3) the most effective way of providing such instruction, e. g., programmed instruction, lecture, small group discussion, etc. Research concerned with these and related problems should provide findings that will prove useful to both teachers and instructional researchers.

FOOTNOTES

1. This article is based upon the Ph.D. dissertation of the first author entitled, "Effect of Training in the Use of Behavioral Objectives and Knowledge of Results on Student Performance in a Mastery Learning Course in Speech Communication" (University Microfilms No. 74-6715). The dissertation was completed under the direction of the second author.
2. This exceptionally low reliability coefficient was attributed to the small amount of variance present in the scores of the six *Ss* completing the test.

REFERENCES

1. Block, J. H., "Operating Procedures for Mastery Learning," in J. H. Block (ed.), *Mastery Learning: Theory and Practice,* Holt, Rinehart and Winston, New York, 1971.
2. Bloom, B. S., "Learning for Mastery," *UCLA-CSEIP Evaluation Comment,* Vol. 1, May 1968.
3. Boardman, D. E., "The Effects of Advance Knowledge of Behavioral Objectives on Students' Achievement in Remedial Chemistry," unpublished doctoral dissertation, University of California at Los Angeles, 1970.
4. Brown, J. L., "The Effects of Revealing Instructional Objectives on the Learning of Political Concepts and Attitudes in Two Role-Playing Games," unpublished doctoral dissertation, University of California at Los Angeles, 1970.
5. Carroll, J. D., "A Model for School Learning," *Teachers College Record,* 4:723-733, May 1963.
6. Duchastel, P. C.; and Merrill, P. F., "The Effects of Behavioral Objectives on Learning: A Review of Empirical Studies," *Review of Educational Research*, 43: 53-69, Winter 1973.
7. Ebel, R. L., "Estimation of the Reliability of Ratings," *Psychometrika*, 16:407-424, 1951.
8. Gergen, K., *Concept of Self,* Holt, Rinehart and Winston, New York, 1971.
9. Jenkins, J. R.; and Deno, S. L., "Influence of Knowledge and Type of Objectives on Subject-Matter Learning," *Journal of Educational Psychology*, 62: 67-70, 1971.
10. Livingston, S. A., "Criterion-Referenced Applications of Classical Test Theory," *Journal of Educational Measurement,* 9: 13-27, 1972.
11. Mager, R. F., *Preparing Instructional Objectives,* Fearon, Palo Alto, 1962.
12. Morse, J. A.; and Tillman, M. H., "Effects on Achievement of Possession of Behavioral Objectives and Training Concerning Their Use," paper presented at the annual meeting of the American Educational Research Association, Chicago, 1972.
13. Tiemann, P. W., "Student Use of Behaviorally Stated Objectives to Augment Conventional and Programmed Revisions of Televised College Economics Lectures," paper presented at the annual meeting of the American Educational Research Association, Chicago, 1968.

THE TYPOLOGY MODEL RE-EXAMINED

JOAN L. GREEN
University of San Francisco

JAMES C. STONE
University of California, Berkeley

ABSTRACT

This report describes the investigators' experiences in the use of cluster and object-analysis techniques (BC-Tryon System) to test the variations in membership and curricular preferences of subgroups of students within a given class over a three-year period. Students' scores on *Q*-sort items describing preferences for curriculum learning experiences, teaching methods and styles, and student-faculty relationships were determined each year and submitted to pre-set cluster analysis. A typology of subgroups, reflecting variations in the cluster scores, was then formed. It was found that given the same population of students each year, there were variations in their individual curricular preferences from year to year and that the membership of subgroups also differed. These findings led to modification of a previously proposed typology model for individualized group learning/teaching.

THE IDEA OF curriculum planning and implementation based on the identified curricular preferences of students has been explored by the authors in two research efforts and several publications (1, 2, 3, 4). A model for curriculum planning in higher education, based on the evaluative perceptions of students, was defined as the stu-

dent typology approach to curriculum implementation. The purposes of this report are to (1) discuss further developments resulting from statistical experimentation with the model, and (2) present their implications for curriculum development.

The Problem

The typology model is an alternative to curriculum implementation which involves simultaneous approaches to the teaching-learning process for a sequence of articulated courses within a subject matter field, such as might be found in professional programs in nursing, engineering, or teaching, or any liberal arts major. The method suggests a way in which students' curricular preferences for various teaching-learning styles can be identified early and used as a basis for instructional planning. In the typology model the individual perceptions of students for the most preferable and least preferable features of the course sequence are obtained through the administration of a *Q*-sort (5). The items of the *Q*-sort are written to describe the more common features of the curriculum under the direct control of the faculty, and reflect its philosophy and objectives. The items within the *Q*-sort are arranged in categories, each of which represents a broader conceptualization or generalization of the program areas which the faculty can adjust or modify according to the curricular preferences of students and yet still achieve course objectives. These categories, for example, might be (1) program aims and objectives; (2) learning experiences; (3) teaching methods; and (4) nature of student-faculty relationships. The result of the students' scores on the individual *Q*-sort items is translated into a composite score for each of the categories, resulting in a set of individual mean category or "cluster" scores for each student (6).

By submitting students' individual cluster scores to further analysis, a typology of subgroups of students may be developed. Each student member of a subgroup manifests a profile of cluster scores similar to the other members of the same subgroup, but different from the members of other subgroups. In other words, members of the same subgroup express the same priorities for features of the course sequence which they want emphasized as well as those which are of least interest, concern, or value to them. The more alike the cluster score profiles of the subgroup members, the greater will be the homogeneity within subgroups and the greater the difference *between* subgroups. Definition of the nature, character, composition, and preferences of each of the subgroups permits personalized curriculum planning designed to capitalize on the preferences of students within the same subgroup. Such program planning optimizes the attainment of the general objectives of the course sequence for each subgroup.

The typology model is based on the following assumptions:

1. It is legitimate to plan the implementation of curriculums upon the preferences of students currently enrolled in a program.
2. Any given group of students is composed of individuals who possess varying preferences and priorities for the various features of the program and can be so differentiated.
3. Curriculum planning and implementation based on the identified preferences of students maximizes the achievement of the aims of the program by all students and increases the likelihood of higher levels of student satisfaction.
4. Curriculum planning and implementation based on the identified preferences of students provides for stimulation, creativity, and flexibility, thereby resulting in greater satisfaction on the part of the faculty.

On the basis of the authors' previous investigations, it was speculated that priorities of students identified early in a course sequence might be used to group students on the basis of their curriculum preferences throughout the remainder of the course sequence [See Figure 1, as originally published in (2).]

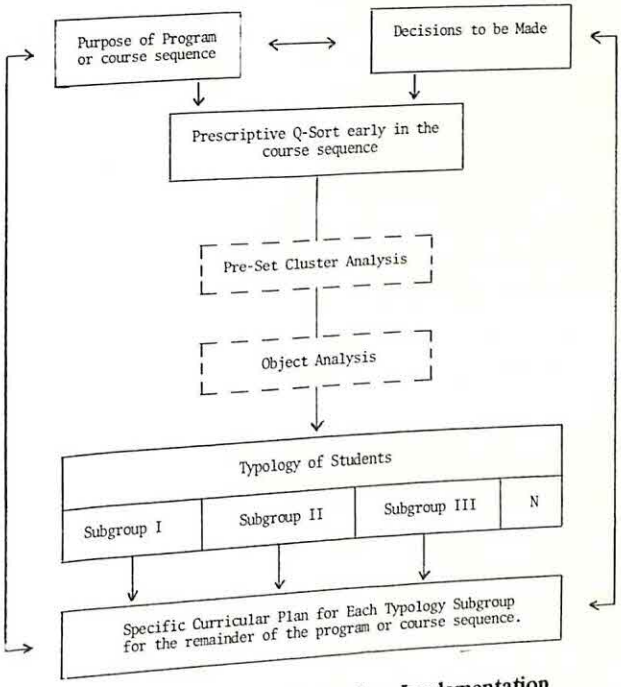


Figure 1.–Typology Model for Curriculum Implementation

Questions to Be Investigated

The specific questions to be answered are:

1. Will a given group of students maintain the same typology structure throughout the period of time they are enrolled in a program or course sequence?
2. Will the membership of a subgroup persist throughout the program or course sequence even though students' curricular preferences may change?
3. Will a subgroup of students change similarly?

The speculation that priorities of students identified early in a curricular experience would persist throughout the remainder of a program (thus facilitating group-oriented curricular planning) is based on observations made during a five-year curriculum evaluation project conducted at the University of San Francisco School of Nursing (2).

In that study it was found that there was a tendency for curricular preferences to persist from year to year. This finding was based on cluster analysis of the students' scoring of *Q*-sort items. Some of the same clusters, describing a given class's curricular preferences (i. e., Class of 1970), tended to reappear from year to year for a given class at each of the sophomore, junior, and senior levels. Other clusters emerged from the data analysis for each level of the curriculum (i. e., sophomore, junior, or senior), irrespective of the group of students, and were believed to be unique to a given level of the program since they represented the learning experiences characteristic of that level of the program.

Thus, some clusters describing curricular preferences seemed to be associated with the "personality" of a given class, i. e., the Class of 1970, while other clusters seemed to be associated with the nature and character of a given level of the program. Students would assess the junior year of the curriculum in two ways: (1) simply because they were junior students evaluating the junior level of the program, and (2) simply because they were either the Class of 1970, 1971, or 1972. It was these findings that led to the development of the typology model and the refinement of the cluster technique so that data collected might be used for both curriculum implementation and curriculum evaluation.

Rational Clusters

The commitment to the development of a model for both curriculum implementation and evaluation led to the use of rational clusters. The rational approach provided for the inclusion of all *Q*-sort items in the clusters. This approach provides a basis from which to evaluate the total program or course sequence since the faculty can use a common base for yearly review of program activities from the same frame of reference. This approach also appears to be a solid base from which to plan for personalized group instruction since the faculty can pre-plan alternatives to instruction and thus need only "plug in" the "right student subgroup" for each option once the students' preferences have been identified.

The authors were curious, therefore, whether a typology of students based on rational clusters also would persist from year to year, or if it would be necessary to define the typology each year that a given group of students was enrolled. Since complete *Q*-sort data (from each of the sophomore, junior, and senior years) were available on two separate classes of University of San Francisco (USF) students (the Class of 1971 and the Class of 1972), those data were analyzed further in order to test the hypothesis that curricular preferences of students did indeed prevail from year to year when using the rational clusters as the basis for organization. It was expected that either acceptance of or failure to reject the hypothesis would lead to further refinement of the typology model and increase its usefulness and generalizability for other educational programs. It was recognized that an important limitation to the results of this analysis would be the fact that the findings would be based on the retrospective perceptions of two classes of students evaluating *common* curricular experiences. The typology model has not yet been subjected to testing through simultaneous approaches to curriculum implementation based on the identified needs of subgroups of students. Such experimentation may well alter again the patterns, membership, and maintenance of typology constructs.

Related Research

Nine clusters were used in the initial research serving as the background for this report. All 72 items of the *Q*-sort designed to obtain student perceptions of the curriculum were assigned to the most appropriate cluster. The clusters were seen as the areas of the USF program over which the faculty had control and were areas in which students would tend to differ in their preferences. They were the generalizable areas of the curriculum and were pertinent to all three levels (sophomore, junior, and senior) of the professional component of the program, independent of specific subject matter content.

Based on the content of the items contained in each of the clusters, two descriptive statements were written to accompany each cluster. The first statement described the kinds of learnings, methods of teaching styles and procedures, evaluation techniques, and/or nature of faculty-student relationships which would be favored as high-priority or highly preferred characteristics of the students who would score high on that cluster. The second statement defined the areas of disinterest or low priority of students who would score low on that cluster. Titles of each of the clusters and the characteristic high-scoring and low-scoring statements describing them follow (1: 51-57):

Cluster I:
Program Objectives Conducive to Professional Attitudes and Understandings (12 items)

High scorers on this cluster are in accord with the value of the professionally oriented objectives of the program and express the highest degree of satisfaction with learning experiences designed especially to effect those goals. Subjects agree that significant curricular experiences should be characterized by learnings which would: (1) help students understand concepts of economics as

they relate to comprehensive and continuous health care; (2) assist students to plan and administer health care based on the fundamental rights and preeminent needs of patients and their families; (3) make students aware of their professional obligation to utilize research findings, effect change, and continue their own education; and (4) develop skills of effective communication and interaction with patients, families, and all levels of health workers.

Low scorers on this cluster are not committed to the professionally oriented objectives of the program and attribute little value to learning experiences stressing the roles and functions of the professional practitioner. Subjects tend to be dissatisfied with the kinds of learning situations valued by high scorers.

Cluster II:
Program Objectives Conducive to the Development of Problem Solving Skills (10 items)

High scorers on this cluster are in accord with the value of program objectives which guide the learning opportunities emphasizing development of professional expertise in resolving health care problems. Subjects agree that significant curricular experiences would be defined by opportunities to gain confidence in their ability to function effectively in all settings with patients of all ages and their families. Subjects agree further that the skills necessary for them to assist others to achieve and maintain high level health require specific laboratory experiences in: (1) establishing therapeutic relationships; (2) recognizing coping measures used in crises; (3) making independent judgments; (4) initiating change; (5) utilizing resource personnel; (6) making referrals; and (7) teaching essentials of health care.

Low scorers on this clustesr are not committed to the value of learning opportunities in the program which emphasize professional problem-solving skills. Subjects tend to place less importance on laboratory experiences requiring them to develop such skill or demonstrate competency in the behaviors valued by the high scorers.

Cluster III:
Arrangement of Learning Opportunities Featured in the Program (8 items)

High scorers on this cluster are in accord with the value of sequential laboratory experiences designed to provide subjects with opportunities to reinforce their learnings through repetition and practice and to move on to more complex learnings as they demonstrate mastery of prior objectives. Subjects agree that settings involving persons from a variety of diverse social classes and cultural backgrounds would provide the most ideal circumstances in which to observe, plan, initiate and administer definitive nursing care, gain skill in performing a variety of technical procedures, and practice leadership functions.

Low scorers on this cluster are less concerned with the arrangement of their learning experiences in an orderly progression and do not agree that diversity in the instructional setting is essential to the mastery of learning objectives.

Cluster IV:
Role of Group Learning and Instruction in the Program (6 items)

High Scorers on this cluster are in accord with the value of didactical methods characterized by the group process. Subjects agree that their learning experiences should feature group projects and conferences, team teaching, and section and seminar meetings so as to benefit from the learnings of their peers and the diverse experiences and interests of the faculty. They concur that group conferences before and after laboratory experiences provide valuable opportunities to communicate learning needs and objectives, clarify theoretical concepts fundamental to nursing practice, and to otherwise prepare for the laboratory itself.

Low scorers on this cluster doubt that group learning and teaching can expedite their own achievement and question the contribution of instruction employing these methods.

Cluster V:
Characteristics of Exemplary Faculty Members (8 items)

High scorers on this cluster are in accord with the value of nurse faculty members who can serve as professional role models. Subjects agree that the ideal instructor keeps up with changes in the practice of nursing and is herself a competent practitioner. As educators, subjects value faculty members who respect students as adults and forthcoming professionals, are prudent in their discussion of matters relating to students, initiate opportunities for exchange of ideas between students and the faculty, and air differences of opinion between and among the students and faculty openly and rationally. Subjects agree further that such collegiality is demonstrated by willingness of faculty members to participate in student-initiated social activities.

Low scorers on this cluster doubt that the exemplary roles of faculty members assume a vital function in their professional education. Subjects are skeptical that collegial interactions between the students and faculty have a positive influence on the learning process.

Cluster VI:
Faculty Behaviors Which Organize Instruction (8 items)

High scorers on this cluster are in accord with the value of learning experiences which are planned, structured, and directed by the faculty. Subjects agree that new learn-

ing experiences, whether clinical or theoretical, should reflect a high degree of faculty intervention to identify the general and specific needs and interests of students and to plan accordingly.

Low scorers on this cluster minimize the necessity for the faculty to organize and intervene in their learning activities or to determine subjects' levels of readiness for new experiences. These subjects do not rely on faculty members to identify or assist in the transfer of their learnings nor are they concerned that the faculty heed student opinions about curricular experiences.

Cluster VII:
Faculty Behaviors Which Individualize Instruction
(7 items)

High scorers on this cluster are in accord with the value of formulating their own learning objectives and selecting the most appropriate experiences by which to achieve them. Subjects agree that in recognizing and honoring student-selected objectives and experiences, the faculty should assist students to choose realistic plans capable of achieving success while simultaneously encouraging alternative approaches and supporting student decisions which might be contrary to those faculty members would make in similar situations.

Low scorers on this cluster do not recommend student-selected or highly individualized learning experiences. They reject the notion of alternate approaches to solving nursing problems and agree that the faculty rather than the students should take the initiative in planning a laboratory experience.

Cluster VIII:
Faculty Behaviors Which Evaluate Learning and Progress
(6 items)

High scorers on this cluster are in accord with the value of evaluation procedures which clarify learning needs and formulate new objectives. Subjects agree that the evaluation process should be individualized in terms of the abilities, interests, and previous experiences of the students and reflect both students' own self-appraisals and the faculty's consideration of external factors influencing student learning and progress. Subjects agree further than in evaluating student achievement, faculty members should expect no more of students than they would of themselves in similar situations.

Low scorers on this cluster doubt that evaluation plays a prominent role in their learning. They question the likelihood that faculty members individualize the evaluation process. Subjects do not agree that the faculty should consider students' self-evaluation nor that evaluation conferences are used to point out areas for improvement.

Cluster IX:
Faculty Behaviors Which Support and Encourage Students
(7 items)

High scorers on this cluster are in accord with the importance of an enthusiastic faculty who make learning come alive for students. Subjects agree that such faculty are readily available to students for consultation and support and manifest interest in students as individuals by: (1) being sensitive to needs for repetition and reinforcement of learning; (2) providing positive feedback related to student progress and achievement; and (3) communicating empathy for the problems encountered in learning. Subjects agree further that supportive faculty members are discreet in their concern for the personal difficulties of students and able to make appropriate referrals for assistance with those difficulties both tactfully and helpfully.

Low scorers on this cluster are less demanding in their expectations of the faculty to provide support and encouragement. They do not value the ability of faculty members to identify with the learning problems of students or to be aware of the personal difficulties of students which influence learning and thus require assistance. These subjects are independent of the need for reassurance regarding their achievements in the nursing program.

Methodology and Findings

Using the object-analysis procedures described in earlier reports, typologies were constructed for each of the two classes at all three levels (1:41-42). Examination of the findings revealed that there *were* differences in the typologies created for the Classes of 1971 and 1972, and that there were differences within the same class at each

Table 1.—Mean Cluster Scores and Standard Deviations for Curricular Preference Typology Subgroups I and II, Class of 1972 as Sophomores

	Subgroup I N = 12		Subgroup II N = 37	
Cluster	Mean	S.D.	Mean	S.D.
1	40.94	9.11	52.61	6.86
2	38.18	5.53	53.44	7.55
3	47.27	7.20	50.71	8.24
4	62.44	6.13	48.76	8.04
5	51.30	7.34	46.47	7.54
6	54.46	8.24	49.14	8.66
7	47.42	4.08	50.19	6.82
8	55.32	4.80	50.08	7.64
9	60.00	7.59	46.56	7.94

of the three levels. The findings for the Class of 1972 are reported in this article. The mean cluster scores and standard deviations for the curricular preference typology subgroups for the Class of 1972 as sophomores are presented in Table 1, for the Class of 1972 as juniors in Table 2, and for the Class of 1972 as seniors in Table 3.

Table 2.—Mean Cluster Scores and Standard Deviations for Curricular Preference Typology Subgroups I-V, Class of 1972 as Juniors

	Subgroup I N = 7		Subgroup II N = 16		Subgroup III N = 18		Subgroup IV N = 12		Subgroup V N = 15	
Cluster	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
1	36.05	6.87	44.35	7.67	46.58	5.64	56.79	8.20	57.87	6.10
2	38.49	6.41	54.16	8.27	43.86	6.43	52.72	5.51	57.22	6.82
3	46.59	3.90	51.14	6.42	42.67	11.63	58.24	7.27	50.69	8.05
4	47.36	8.77	47.20	8.14	54.95	7.30	59.18	8.15	40.27	4.71
5	58.20	8.85	45.46	5.26	55.40	8.67	45.68	8.83	50.51	7.78
6	66.14	3.77	49.92	8.19	54.73	4.61	44.91	8.18	44.50	5.51
7	52.25	7.06	60.73	5.65	46.18	5.55	42.71	7.59	43.43	6.91
8	61.96	5.95	50.69	5.47	55.31	8.31	37.32	3.75	47.75	6.25
9	60.26	5.26	49.93	8.03	54.12	5.86	43.87	8.09	48.14	8.23

Table 3.—Mean Cluster Scores and Standard Deviations for Curricular Preference Typology Subgroups I-V, Class of 1972 as Seniors

	Subgroup I N = 14		Subgroup II N = 10		Subgroup III N = 21		Subgroup IV N = 15		Subgroup V N = 8	
Cluster	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
1	61.83	5.40	40.10	7.97	51.06	7.83	50.80	6.05	47.86	6.71
2	63.02	3.69	42.63	7.35	50.89	6.52	45.43	6.69	54.20	8.09
3	49.25	7.95	47.40	4.46	52.47	5.78	41.13	8.08	62.99	7.19
4	47.15	6.76	41.08	6.36	54.57	6.72	51.11	7.83	39.40	3.16
5	40.78	7.74	45.09	8.59	52.92	6.83	55.11	5.65	49.75	5.74
6	44.96	7.93	56.90	6.22	47.96	8.01	50.15	7.30	42.21	8.00
7	48.83	9.13	54.80	6.87	42.82	8.67	60.05	5.23	48.26	7.93
8	39.14	4.37	65.21	8.82	47.06	7.30	54.78	6.39	46.20	4.37
9	42.84	8.55	57.99	5.69	49.57	6.93	46.17	6.61	49.10	6.61

At the sophomore level, the Class of 1972 was arranged into two subgroups of similar profiles of curricular preferences, and into five subgroups at each of the junior and senior levels. In each instance, a number of students were unassigned to the emerging subgroups because their profiles of curricular preferences were not only atypical of the classifiable subgroups, but also were atypical of each other's preferences. Curriculum planning for "atypical" students would have to be handled on a separate basis in the typology model.

When the nine clusters are arranged in categories according to their mean cluster scores for each of the subgroups at each of the three levels, it can be seen that the priorities of the subgroups within a class tend to differ (Tables 4, 5, and 6). Though there are five subgroups formed within the population at each of the junior and senior levels, they are not necessarily the same subgroups, nor will the membership assigned to those subgroups be constant.

The very high (or high) and very low standardized mean cluster scores (between 40-45 and over 60) are generally interpreted as the defining characteristics of the curricular preferences for each of the typology subgroups. In other words, the extreme mean cluster scores define the highest and lowest priorities of a given subgroup for their curricular preferences. The cut-off points may well vary depending on the range and distribution of the mean cluster scores for the members of a given subgroup.

When the pertinent clusters defining the priorities of a subgroup for the teaching-learning process have been identified, the appropriate high- or low-scoring statements describing curricular preferences are then reviewed for their curricular implications. For example, at the junior level (Table 5), a composite statement could be written combining the high-scoring statements accompanying Clusters 6, 8, and 9 and the low-scoring statements for each of Clusters 1 and 2. The resulting profile would be a prescription to the faculty for the kinds of learnings most preferred and most likely to be avoided by the members of that subgroup. More explicitly, junior students in Subgroup 1 favored organized instruction (Cluster 6), faculty control and guidance over evaluation of learning activities (Cluster 8), and much support and encouragement from the faculty (Cluster 9). These subgroup members were also less concerned about the professional aims, objectives, and content of the program (Cluster 1) or the development of problem solving skills (Cluster 2). Certainly it would be helpful for the faculty to know who these students are—in advance of semester planning—so as to plan better for the achievement of the program objectives by those students and to adopt the most useful teaching methods.

In the senior year (Table 6), it can be seen that the second subgroup of the typology was similar to Subgroup I in the junior year typology of the same class. Particularly interesting was the discovery that the membership of Subgroup I in the junior year typology for the Class of 1972 was not the same as the membership of Subgroup II in the senior year typology for the same Class of 1972. Though data concerning the group memberships are not reported in this article, it can be seen that it is somewhat fallacious to depend on stereotypes about given students for curriculum planning since they may well change. On the other hand, it is equally specious to wait all year to find out what a student's needs and preferences are. If means can be found to identify students' curricular preferences and preferred learning and teaching styles early at the beginning of the school year, why not do so—and plan accordingly?

If a typology structure can be obtained for a given class of students at the end of the previous year's learning experiences, or very early at the beginning of the current year's learning experiences, composite statements describing each of the subgroups' preferences for teaching-learning processes as normally presented in the curriculum can be used by the faculty to:

1. individualize the curriculum for each subgroup;
2. match the appropriate teacher(s) with each subgroup;
3. choose, deliberately and consciously, the learning theories and teaching strategies most appropriate for assisting each subgroup of students to achieve the aims of the program.

It is essential to understand that the aim of the typology model is to facilitate the achievement of the common aims or goals of a structured program through the use of a variety of instructional methods and procedures which are based on (or chosen on the basis of) the identified preferences of subgroups of students within the same class. In other words, it should not be presumed that a faculty would always work with, or choose to reinforce, the highest priorities of a given group of students. It may well be in the better interests of the students to attempt to alter or change the students' priorities so as to assist them to meet the objectives of the program. The point is that with the typology model, the instructional decisions (application of theory, use of teaching strategy, selection of learning experience, etc.) are made knowledgeably by faculty well informed about the characteristics of their students. Utilization of the model by a faculty assumes the role of group learning and teaching in the program and presupposes that the faculty members are committed to the philosophy of individualized instruction. Use of the model provides a means for combining the major advantages of both group learning and individualized instruction.

Implications

On the basis of the retrospective analysis of typology formation at three levels of the curriculum for two separate classes of students, it appears that the typology model needs to be modified. Rather than determining the typology structure once, and planning accordingly for each of the ensuing years in the program, it now appears that it would be desirable to redefine the typology

Table 4.–Nine Rationally Defined Clusters Arranged in Categories According to Mean Cluster Scores (MCS) for CPT Subgroups I and II, Class of 1972 as Sophomores.

Subgroup Number	Very Low (MCS 40-45)	Med. Low (MCS 45-50)	Med. High (MCS 50-55)	High (MCS 55-60)	Very High (MCS 60+)
I	2* 1	3 7	5 6	8 9	4
II		5 9 4 6	8 7 3 1 2		

*MCS below 40.00

Table 5.–Nine Rationally Defined Clusters Arranged in Categories According to Mean Cluster Scores (MCS) for CPT Subgroups I-V, Class of 1972 as Juniors.

Subgroup Number	Very Low (MCS 40-45)	Med. Low (MCS 45-50)	Med. High (MCS 50-55)	High (MCS 55-60)	Very High (MCS 60+)
I	1* 2*	3 4	7	5	9 8 6
II	1	5 4 6 9	8 3 2		7
III	3 2	7 1	9 6 4	8 5	
IV	8* 7 9 6	5	2	1 3 4	
V	4 7 6	8 9	5 3	2 1	

*MCS below 40.00

Table 6.–Nine Rationally Defined Clusters Arranged in Categories According to Mean Cluster Scores (MCS) for CPT Subgroups I-V, Class of 1972 as Seniors

Subgroup Number	Very Low (MCS 40-45)	Med. Low (MCS 45-50)	Med. High (MCS 50-55)	High (MCS 55-60)	Very High (MCS 60+)
I	8* 5 9 6	4 7 3			1 2
II	1 4 2	5 3	7	6 9	8
III	7	8 6 9	2 1 3 5 4		
IV	3	2 9	6 1 4 8	5	7
V	4* 6	8 1 7 9 5	2		3

*MCS below 40.00

structure each year or at each level of a course sequence (Figure 2). Even this feature is subject to change, depending on the impact of individualized group-oriented curricular approaches, It well may be that once the typology approach is instituted, the typology structure and the membership of the typology subgroups *would* remain constant from year to year. Experimentation in this area is recommended to test the model further.

If the typology is to be developed at each level of the curriculum, the way is open for the faculty to develop a *Q*-sort unique to the learning experiences for each level of the program or course sequence. The items, however, still should reflect the philosophy, aims, and program objectives generalizable to all curricular areas. The *Q*-sort items may reflect options or alternative means to achieving certain combinations of objectives for a given level of the program. Development of the typology prior to instituting the program will assist the faculty in the proper match of students to the learning experiences and teaching styles available in the program.

The typology might be constructed on the basis of an instrument common to all levels of the program or on the basis of an instrument unique to each level. Further, the typology might be developed at the end of one level of the program for planning for the next level, or at the beginning of the academic year at each level. These decisions are inherent in the nature of the *Q*-sort instrument developed for use, the amount of pre-planning done in terms of curriculum options, and the faculty's beliefs concerning the amount of change which might occur in a given group's preferences between the end of one level of the program and the beginning of the next, and the nature of the course sequence or program.

The usefulness of typology construction at the end of the program might be questioned. In the authors' viewpoint this procedure is essential to the acquisition of baseline data necessary for follow-up studies of program graduates and for continuous ongoing program evaluation. The use of the same clusters at each level of the program from year to year would provide a common base for the formative evaluation necessary for interim feedback and program improvement.

The authors' experience with the typology model raises the question; "Why build a battleship if a canoe

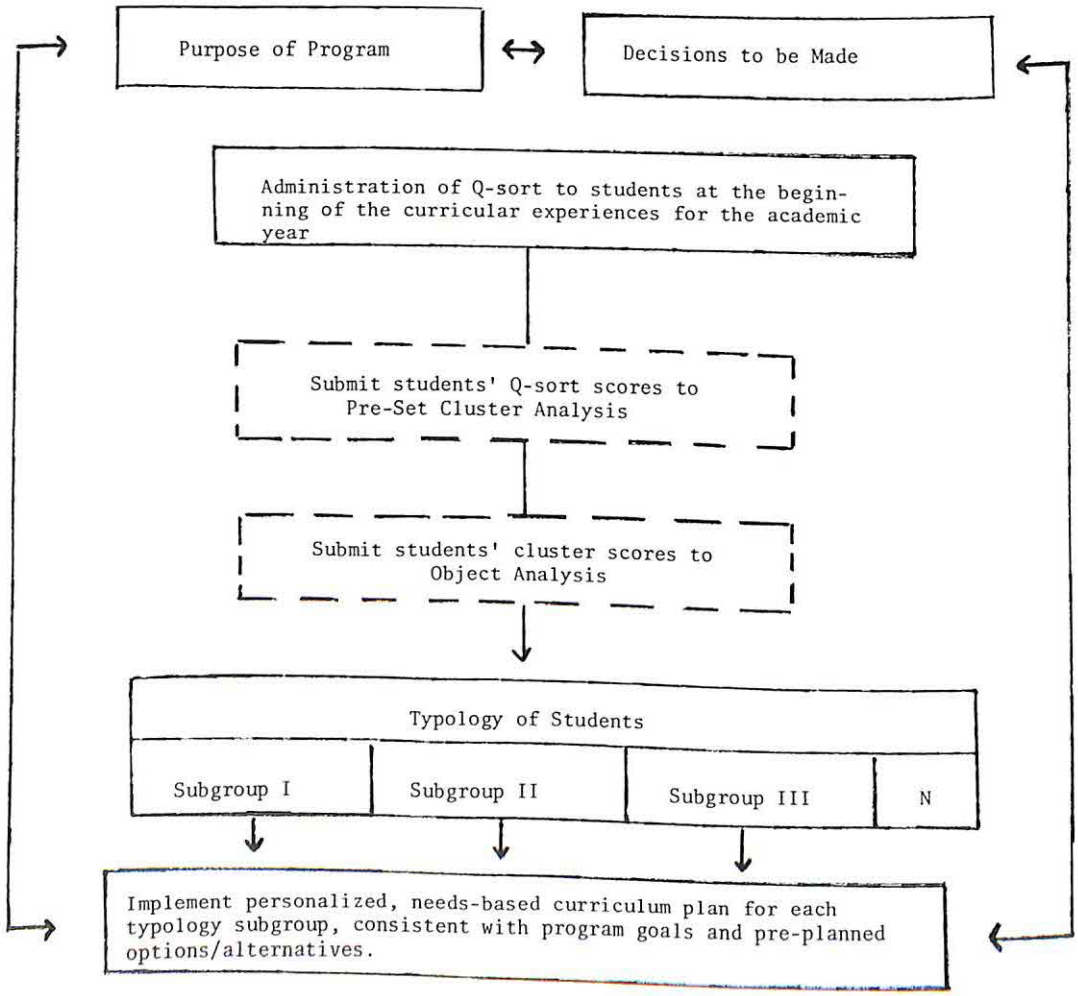


Figure 2.–The Revised Typology Model for Curriculum Implementation

will get you where you are going?" Why go to the extreme of developing a typology when in fact one could find the answer to students' preferences by just asking them? The authors are not convinced that exotic methods in themselves are more productive or effective than simpler techniques. But do faculty members consider the question of student needs, characteristics, and preferences, let alone ask it, when curriculum planning? There is merit in developing a systematic approach to planning and implementing curriculums. Turnover of faculty alone is a major consideration. Is it not more effective for a patient's history and record of diagnostic procedures to travel with him to a new physician? We deplore submitting patients to needless or repetitious procedures. Why do we submit our students to constant, and unsystematic, procedures in defining their learning needs and preferences? More explicity, why do we persist in assuming that there are no differences among our students? Not only is the typology method an efficient, controlled, and fairly simple-to-implement strategy to diagnose learning needs of students; it also provides the assurance of validity for those faculty members who find comfort in statistical verification rather than subjective hunches about the needs of students. Furthermore, the typology model provides a method for creative curriculum planning which minimizes the likelihood of unplanned, unwanted, or unseen changes in the curriculum. The authors contend that use of the typology model provides a means for making faculty members more conscious of the teaching-learning process and the needs of the students whom they instruct.

Summary

The typology model, which differentiates the preferences of students, is an aid in fostering the diversity of students in higher education. When preferences of students have been identified, the faculty can work with them, adopt definitive teaching strategies, and formulate learning objectives and experiences to enhance those differences. Use of the typology model provides a mechanism which facilitates implementation of programs based on mastery learning. Individualized learning, through the group structure, may shorten the amount of time needed for a student to achieve the aims of a program.

The need for conscious structuring of curricular approaches to implementation of professional education programs is enhanced by the heavy demands made by society and the professions to clarify levels of professional practice, formulate new definitions and standards of professional practice, and to accelerate passage through the career ladder. The authors believe that the many contingencies facing professional educators in academic and clinical settings deter faculty from a conscious awareness of their roles as teachers and of teaching and learning as processes. While individualized curricular approaches to professional education will not solve all the problems faced by the educator, it is hoped that the typology approach to curriculum implementation will single out the primary focus of the educator's role—the student—and provide a means by which faculty members can reorder their priorities to provide greater emphasis on meeting students' needs.

REFERENCES

1. Green, Joan L., "The Relationship between Membership in a Curricular Preference Typology and Selected Performance Outcomes," unpublished doctoral dissertation, University of California, Berkeley, 1974.
2. Green, Joan L.; and Stone, James C., "Teach Me and I Will Be Silent," *Development and Use of Tools for Curriculum Evaluation*, Report of the Five-year Study, USPHS Nurse Project Training Grant No. D 10 NU 00235, 1968-1973, University of San Francisco, San Francisco (Available from ERIC Clearinghouse, Washington, D. C., 1974.)
3. Green, Joan L.; and Stone, James C., "Models for Curriculum Evaluation in High Education," *Educational Record* (in press).
4. Green, Joan L.; and Stone, James C., "A Student Typology Model for Curriculum Implementation," *Nursing Forum* (in press).
5. Stone, James C.; and Green, Joan L., "The Double *Q*-Sort as a Research Tool," *Journal of Experimental Education*, 40:81-88, Fall 1971.
6. Tryon, Robert C.; and Bailey, Daniel E., *Cluster Analysis*, McGraw-Hill, New York, 1970, pp. 1-5.

THE EDUCATIONAL FORCES INVENTORY: A NEW TECHNIQUE FOR MEASURING INFLUENCES ON THE CLASSROOM[1]

NICHOLAS F. RAYDER
BART BÖDY
Far West Laboratory for Educational Research and Development
San Francisco, California

ABSTRACT

The Educational Forces Inventory (EFI) is a technique for assessing the constellation of forces in the teacher's social-psychological field. Thereby, important stress points for teachers are identified and their work in the classroom is facilitated. The EFI charts the relative influence of each of 13 forces that are important factors in the educational settting. Teachers are asked to characterize each force along two dimensions: the amount of influence it exerts upon the classroom process, and the degree to which this influence is positive or negative. Several hundred teachers and teaching assistants in Follow Through classrooms throughout the country participated in a field test of the instrument, with concurrent administration of the Purdue Teacher Opinionaire (PTO), an established measure of teacher morale. The high rate of return and correct completion on the EFI indicated that it is a practical technique. In addition, the data indicated validity for the instrument in three important respects: (1) the pattern-of-importance ratings correspond to independently assessed patterns of physical and social distance; (2) the positive/negative ratings for particular forces correlated substantially with corresponding subscores of the PTO; and (3) the pattern of responses reflected aspects of internal consistency. Some ways of utilizing EFI data are discussed: how to plot and interpret two-dimensional force fields, and how to use them in program implementation, especially to facilitate the work of the teacher in the classroom.

SITUATIONAL VARIABLES OF the educational setting have a significant influence on teachers and on the teaching/learning process in the classroom. Little is known about the extent or the type of influence of specific variables. Dreeban (4) has stated, "The study of the impact of the environment, both within school systems and from the external community, on the work of teachers has barely begun."

In mapping the relationship of teachers to the environmental setting, the usual approach is to focus on the characteristics of teachers and their attitudes. For instance, the Minnesota Teacher Attitude Inventory (3) and the Purdue Teacher Opinionaire (2) both refer to "teacher morale" as the central variable. But if we are to improve the lot and the effectiveness of teachers, we must go beyond teacher attitudes, and try to identify the salient situational elements which affect both teacher morale and teacher effectiveness.

Here the authors present a technique which charts the educational setting considered as the teacher's work-space. The name of this technique is the Educational Forces Inventory, or EFI. It has been successfully tested and applied "in the field." In this paper data are presented on the initial validation of this technique, and a discussion of how to plot force fields, how to interpret them, and how they can be made useful for program implementation and for modifying classroom processes is included.

The concept of a psychological field of force was developed by Lewin (5) as a means of understanding an individual's behavior in relation to his environment—to provide a common frame of reference for the interplay of "internal states" and of "objective reality." For instance, a child that is repeatedly prevented from approaching a desired object, say a ball, eventually erects an internal barrier which allows him to "forget" about it and thus avoid further frustration. A force field reflecting this situation will show the child in relation to the desired object, with an intervening barrier making it inaccessible. In addition, this representation may depict a whole range of other dynamic consequences of the barrier, such as a generalized constriction of the field that may inhibit the child's locomotion in ways otherwise unrelated to the ball, and persisting long after the ball has gone. In short, a force field is an "open" system for taking into account, at a somewhat abstract level, any number of facts, such as circumstances in the environment and patterns of behavior.

The concept of field of force was later extended to deal with sociological phenomena (6). A different type of

force field, called a "phase space," was used to chart characteristics of a group—such as ethnic prejudice or rate of production in a factory—as functions of a multiplicity of forces, acting over time:

> Food habits of a group, as well as such phenomena as the speed of production in a factory, are the result of a multitude of forces. Some forces support each other, some oppose each other. Some are driving forces, others restraining forces. Like the velocity of a river, the actual conduct of a group depends upon the level (for instance the speed of production) at which these conflicting forces reach a state of equilibrium (6).

The impetus for applying force field analysis in the context of social groups was related to a series of studies which demonstrated that efforts directed at individuals were relatively ineffective in changing social behavior, as compared to the effects of group process; and that often change achieved in an individual context was short-lived, at best, if not accompanied by corresponding changes in group standards (1, 6, 8).

To effect lasting institutional change it is necessary to deal with the institutional environment itself. The logical way to begin is to chart this environment as a field of forces, each with more or less power, and each more or less directed toward or away from the desired state of affairs. The EFI has been specifically designed to accomplish this purpose in the educational milieu. The focal variable, teacher effectiveness or classroom process, is a joint characteristic of individuals and of educational settting; the "window" through which it is viewed is teacher attitudes; and the forces that are charted as more or less powerful and more or less helpful are easily identifiable entities within the teacher's work-space.

Development and Initial Validation of the EFI

Designing the EFI: Background and Objectives

The Responsive Educational Program, sponsored by Far West Laboratory, is designed to improve the form of educational experience offered the child, especially with respect to classroom process (9). It is therefore concerned with improving the work-space and the effectiveness of the teachers by bringing about far-reaching changes in the educational setting and by providing inservice training. Implementing this model within local school districts, in the context of the national Follow Through program,[2] involved some fundamental changes in the interrelationships of school and community, of students, teachers, and parents, and of teachers, teaching assistants, and teacher-trainers. An important program evaluation objective, therefore, was to determine how participating teachers perceived those characteristic features of the program which have implications for their social-psychological field of force.

To provide a description of the teacher's social-psychological field, a technique must meet several practical requirements. First of all, teachers will understand its use and accept it. Second, teacher responses will reflect independently verifiable facts in the external world, as opposed to personal or role-oriented behavior. Third, the elements will be differentiated into some meaningful patterns, conducive to making decisions about program implementation.

Elements of the EFI

In delineating forces in the teacher's social-psychological field, we sought to identify prominent elements in the environment in the same terms as they overtly and tangibly present themselves to the teacher. On the basis of previous experience in implementing inservice training programs, the following ten forces were selected as important in influencing teacher morale and classroom effectiveness in public schools generally:

1. *Principal* of the school
2. *Central Office* administrative personnel
3. *Other Teachers* in the school
4. *Parents* of children in the class
5. *Curriculum* prescribed by the district
6. *Testing* programs
7. *Statewide Mandates* on certification, curriculum, grading, etc.
8. *Physical Facilities* available
9. *Social Environment* of the community
10. *Curriculum Personnel* such as reading specialist, art teacher, etc.

In addition there were three program components of particular relevance to implementation of the Responsive Educational Program in the context of Follow Through:

11. *Program Director*—coordinator of the Follow Through Program within the district: responsible for administration, community organization, and policy matters.
12. *Program Advisor*—delivers the program to the classroom with inservice training and in-class assistance: each advisor responsible for about ten classrooms.
13. *Other Adult* in the classroom—teacher or teaching assistant: a teaching assistant in this model is a full-time, paid paraprofessional assigned to the classroom and engaged in teaching activities.

Recording Teachers' Attitudes on the EFI

To reflect the constellation of forces in the social-psychological field, it is necessary to identify not only the forces to be assessed, but also the dimensions along which they are to be measured. Given any field, at least two dimensions would be required for an adequate specification. Consequently, the effects of individual forces upon processes in the classroom were to be specified in terms of both power—or amount of influence—and affect—or the degree to which influence is positive or negative in direction.

Table 1.—Means and Standard Deviations of Scores Assigned by Teachers and Teaching Assistants on all Forces and Tasks of the Educational Forces Inventory

Forces	Teachers (N = 214)						Teaching Assistants (N = 180)					
	Task 1[1]		Task 2[2]		Task 3[3]		Task 1[1]		Task 2[2]		Task 3[3]	
	$\bar{X}$	sd	$\bar{X}$	sd	$\bar{X}$	sd	$\bar{X}$	sd	$\bar{X}$	sd	$\bar{X}$	sd
1. Principal	4.65	3.04	11.71	10.29	1.87	1.01	4.84	3.53	12.08	10.12	1.82	.99
2. Central Office	10.11	2.97	2.51	4.23	2.61	.77	8.78	3.76	4.12	6.95	2.65	.90
3. Other Teachers	7.33	3.51	5.96	6.74	2.19	.91	8.05	3.30	4.19	4.86	2.63	.91
4. Parents	6.57	2.90	8.30	6.54	2.29	.98	7.18	3.36	6.63	7.77	2.51	1.01
5. Curriculum	5.08	3.41	11.09	12.51	2.09	.90	6.27	3.30	6.41	7.73	2.33	.99
6. Testing	9.71	2.95	3.78	7.20	2.94	1.00	8.45	3.23	3.18	5.07	2.76	.96
7. Statewide Mandates	10.53	2.73	1.87	3.05	2.84	.75	9.71	3.04	2.32	3.49	2.87	.94
8. Physical Facilities	5.62	3.25	10.24	10.06	2.18	1.09	7.47	3.34	6.20	8.03	2.51	1.07
9. Social Environment	6.59	3.79	9.08	11.31	2.50	1.07	8.23	3.02	4.15	4.78	2.66	1.00
10. Curriculum Personnel	8.08	2.75	4.32	4.72	2.13	.80	7.59	2.98	4.41	6.18	2.40	.90
11. Program Director	7.11	3.39	6.72	9.07	1.91	.82	5.87	3.56	10.08	11.28	1.88	.95
12. Program Advisor	5.16	3.03	10.34	8.68	1.86	.95	5.17	3.28	11.20	10.78	1.76	.92
13. Other Adult[4]	4.50	3.36	16.77	15.44	1.62	.88	2.95	3.12	28.43	21.42	1.40	.82

1 The lower the number, the higher the rated importance.
2 The higher the number, the greater the rated importance.
3 The lower the number, the more positive the influence.
4 The teacher is rating the teaching assistant and the teaching assistant is rating the teacher.

In completing the instrument, the respondent—teacher or teaching assistant—is asked to evaluate the set of 13 forces by carrying out three successive tasks:

Task 1: The 13 forces are ranked in order of their importance in influencing teaching. The force with the strongest influence, either positive or negative, is given the rank of 1, the least important the rank of 13.

Task 2: Each force is assigned a weight according to its relative importance in influencing teaching. A total of 100 points are distributed among the 13 forces, with the most important force assigned the most points. Any pattern of assignments is permissible: the respondent might choose to distribute the points evenly among the 13 forces or allocate them all to just one or two, with no points to the rest.

Task 3: Each force is rated on a scale of 1 to 5, according to its positive/negative effect on teaching, with a rating of 1 indicating strong positive influence, and a rating of 5 indicating strong negative influence.

Design for Field Testing the EFI

A field test was planned to determine if the EFI was a viable technique, and specifically if it met its objectives and satisfied the criterion related to practical utility. In order to test the clarity and acceptability of the procedure, a trial with a large, unselected group of teachers was planned. To test whether the instrument did, in fact, reflect objectively ascertainable factors in the teacher's workspace, the degree of similarity in the response patterns of respondent pairs would be related to their degree of similarity along important dimensions of the social-psychological field, such as professional role (teacher vs. teaching assistant) and operational unit (classroom, school, district). Third, it was planned to relate responses on the EFI to those on an older, established instrument.

A Concurrent Validity Criterion: The Purdue Teacher Opinionaire

The Purdue Teacher Opinionaire (PTO) was chosen as a referent for assessing concurrent validity of the EFI. This instrument had been developed and validated as a measure of teacher morale (2). While the objectives of the EFI extend far beyond teacher morale, no instrument was found with a scope equally broad. The PTO manual cites several studies which reported that scores of teacher correlated appreciably with those of their principals, as well as with

Table 2.–Rankings of Scores Assigned to Forces on Each of Three Tasks by Teachers (N = 214) and Teaching Assistants (N = 180)

Forces	Ranks for Task 1		Ranks for Task 2		Ranks for Task 3	
	Teachers	Teaching Assistants	Teachers	Teaching Assistants	Teachers	Teaching Assistants
Principal	2	2	2	2	3	3
Central Office	12	12	12	11	11	10
Other Teachers	9	9	9	9	8	9
Parents	6	6	7	5	9	7.5
Curriculum	3	5	3	6	5	5
Testing	11	11	11	12	13	12
Statewide Mandates	13	13	13	13	12	13
Physical Facilities	5	7	5	7	7	7.5
Social Environment	7	10	6	10	10	11
Curriculum Personnel	10	8	10	8	6	6
Program Director	8	4	8	4	4	4
Program Advisor	4	3	4	3	2	2
Other Adult	1	1	1	1	1	1
Spearman Rhos	(.95)		(.85)		(.97)	

other logically related aspects of the educational setting that could be assessed independently, such as teacher turnover.

In its current form, the PTO has 100 items which, in addition to the overall score for morale, yield ten subscores reflecting dimensions identified by factor analysis:

a. Teacher Rapport with Principal
b. Satisfaction with Teaching
c. Rapport among Teachers
d. Teacher Salary
e. Teaching Load
f. Curriculum Issues
g. Teachers' Status
h. Community Support of Education
i. School Facilities and Services
j. Community Pressures

Method and Sample

In the spring of 1972, some 300 teachers and a like number of teaching assistants, in 14 school districts in 12 different states, were working with the Far West Laboratory to implement a Responsive Educational model in Follow Through classrooms, kindergarten through third grade. The on-site program advisors in each district gave a copy of the two instruments, the EFI and the PTO, to each teacher and teaching assistant. The instructions were to complete and return them directly to the Laboratory.

The form for the EFI provided for several items of descriptive information to be filled in by the respondent, such as name, age, role (teacher/teaching assistant), and amount of teaching experience. Strict confidentiality was pledged, and in addition specific allowance was made for the option of leaving off the name.

Results

The forms were received by a total of 572 teachers and teaching assistants. Of these total recipients, 428 (or 75%) correctly completed and returned at least the EFI; both forms were completed and returned by 394 (214 teachers and 180 teaching assistants) of these 428. For teachers, the mean number of years of age reported was 35, and of teaching experience, 9. All but two of the teacher-respondents were women, and all but twenty had previously taught in the Follow Through program.

Forces data were analyzed separately for teachers and for teaching assistants. For each of the 13 forces, the following statistics were computed for both groups: the mean of the ranks assigned in Task 1, the mean number of weight points assigned in Task 2, and the mean ratings on the positive/negative continuum given in Task 3. These means and the corresponding standard deviations are reported in Table 1. These means were then rank-ordered within task: the rank scores are presented in Table 2.

For each force, product-moment correlations were computed between scores assigned on each of the three

Table 3.–Correlations* between Scores on Tasks 1, 2, 3 for Each of the 13 Forces for Teachers (N = 214) and Teaching Assistants (N = 180)

Forces	Tasks 1, 2(rating x points)		1, 3(ranking x pos-neg rating)		2, 3(points x pos-neg rating)	
	Teacher	Teaching Assistant	Teacher	Teaching Assistant	Teacher	Teaching Assistant
1. Principal	63	52	38	54	17	38
2. Central Office	58	41	17	25	03	23
3. Other Teachers	68	40	48	28	32	26
4. Parents	15	46	33	32	08	29
5. Curriculum	61	44	40	27	39	23
6. Testing	37	29	13	22	-05	16
7. Statewide Mandates	51	40	13	06	14	08
8. Physical Facilities	65	54	24	39	17	42
9. Social Environment	69	37	09	10	01	15
10. Curriculum Personnel	45	16	42	20	36	09
11. Program Director	52	41	45	42	25	38
12. Program Advisor	62	46	35	47	17	32
13. Other Adult	47	42	28	25	22	25
Average correlation	(.53)	(.41)	(.29)	(.29)	(.18)	(.25)

*Correlations involving task 2 have been inverted in sign to adjust for variations from task to task in the scheme for assigning numbers: on task 1, high influence received a rank of 1; on task 2, high influence received the most points; on task 3, the most positive received a score of 1.

Table 4.–Intercorrelation between Task 3 of the Educational Forces Inventory (EFI) and the Ten-Factor Subscores of the Purdue Teacher Opinionaire (PTO) Collected on 394 *Follow Through* Teachers and Teaching Assistants

	PTO Ten Factor Sub-Scores									
	Rapport with Principal	Satisfaction Teaching	Rapport Among Teachers	Teacher Salary	Teacher Load	Curriculum Issues	Teachers' Status	Community Support	School Facilities	Community Pressures
EFI - 13 Forces	1	2	3	4	5	6	7	8	9	10
1. Principal	(68)**	27	38	12	22	42	09	12	10	02
2. Central Office	17	15	12	16	07	17	24	17	11	06
3. Other Teachers	20	17	(43)	15	15	24	03	12	05	16
4. Parents	04	11	18	13	01	16	16	(35)	08	15
5. Curriculum	18	10	03	15	14	(26)	12	24	06	-03
6. Testing	-04	09	03	10	19	17	13	08	13	11
7. Statewide Mandates	-03	-04	03	04	08	08	07	07	00	00
8. Physical Facilities	21	26	22	08	09	22	11	12	(20)	03
9. Social Environment	09	17	05	00	-03	11	-02	18	04	05
10. Curriculum Personnel	11	06	12	13	-13	12	07	16	11	03
11. Program Director	30	29	16	-01	18	22	20	20	10	00
12. Program Advisor	28	27	22	-09	19	27	08	17	22	14
13. Other Adult	14	20	20	-07	15	17	16	15	16	22

*All entries have been inverted in sign to compensate for a difference in direction of scores: on the PTO a high score indicates "good" or "high" morale, whereas the EFI defines "1" as the most positive, and "5" as most negative.

**Correlations at the intersection points of corresponding EFI forces/PTO factors have been circled.

different pairings of the three tasks, and are presented in Table 3.

The ten-factor subscores of the PTO defined in the manual were correlated with scores on Task 3 of the EFI for all 13 factors. The correlation matrix is presented in Table 4.

Patterns of Teachers and Teaching Assistants

Table 1 suggests that the power and affect attributes of each force were perceived similarly by teachers and by teaching assistants. Table 2 confirms this. In all three tasks, the mean scores calculated for the 13 factors gave rise to very nearly the same rank-order sequence within each of the two groups. For the three tasks, the Spearman rhos between the two groups are: .95; .85; and .97. Both groups considered *Other Adult*, *Principal*, and *Program Advisor* to be the most important and most positive influences: the mean scores of these three ranked among at least the top four on all three tasks for both groups. *Program Director* and *Curriculum* were also rated both influential and positive. Forces such as *Central Office*, *Testing*, and *Statewide Mandates* were considered to be the least important and the least positive influences.

These data suggest that the various forces are perceived as both more influential and more valued in direct relation to the extent of their physical and psychological proximity to the classroom: the nearer the force is located, the more it is seen as powerful and positive, and the further away, the more it is seen as weak and less positive. This direct correspondence between EFI scores and a salient and logically related aspect of the teacher's social-psychological field offers an indication of construct validity for the instrument. Moreover, these perceptions apply across work-roles: they are as true for a certified teacher as for a paraprofessional teaching assistant. This is another indication of construct validity in that the correspondence between EFI scores and the educational setting is not primarily dependent on respondent characteristics.

Relationships among the Three Tasks

For particular forces, the scores assigned in Task 1 and in Task 2 are correlated moderate-to-high. This is to be expected from the instructions, which ask for different expressions of judgments along the same dimension—relative strength of influence. This finding may be taken to be a reflection of internal consistency.

The correlations between either of the first two tasks and Task 3 are low-to-moderate. This again is consistent with the instructions: Task 3 asks for ratings along a dimension that is distinct from the one involved in the other two tasks—positive/negative valuation rather than relative strength of influence. These low-to-moderate correlations, then, indicate that, psychometrically as well as conceptually, Task 3 represents a dimension that is distinct from the one underlying the first two tasks. While distinct, the two dimensions are by no means orthogonal: in most patterns (though not all), the more a force is seen as having strong influence, the more it is seen as having also a positive direction of influence.

The correlations between Task 3 and Task 2 are lower than those between Task 3 and Task 1. Presumably this is because the scores on Task 2 are more volatile: many respondents assigned all of their 100 points to just one or two forces, leaving zero to the rest.

Relationship of the EFI to the PTO

The PTO includes dimensions that are, from their labels and from their definitions in the manual, similar to forces on the EFI. On the PTO a high score indicates a "good" rating, or high morale, whereas the EFI Task 3 defines a score of 1 as most positive, and a score of 5 as most negative. To adjust for this inversion of scoring direction, the signs on the correlation coefficients in Table 4 have been inverted to make a positive item reflect a positive relationship between corresponding concepts. Positive correlations are evidence of concurrent validity. In fact, moderate-to-high positive correlations occur in three of the five instances where the name-to-name relationship is self-evident:

.68 between (1) *Principal* and (a) Rapport with Principal
.43 between (13) *Teacher/Assistant* and (c) Rapport among Teachers
.35 between (4) *Parents* and (h) Community Support

In the other two instances the correlations are low, but also positive:

.26 between (5) *Curriculum* and (f) Curriculum Issues
.20 between (8) *Physical Facilities* and (i) School Facilities and Services

This series of correspondences between EFI forces and subscales of the PTO offers clear evidence of concurrent validity; since the conceptual match between corresponding elements is only approximate, we did not expect uniformly high correlations.

Implications for Program Implementation

One way to apply these data for evaluating program implementation is to look at the influence exerted by the three forces that represent the program. All three are important for delivering program concepts and techniques to the classroom. The presence of the *Teaching Assistant*, if properly utilized, improves the teacher/pupil ratio by a factor of two, making the child's classroom experience more responsive to him. The *Program Advisor* works directly with teachers to translate program goals and concepts into classroom process. The *Program Director* provides administrative and social support, especially to the teaching assistant, whose very role is made possible by Follow Through.

The pattern of EFI scores indicates that the impact of Follow Through on classroom teaching was both powerful

and positive. Table 2 presents the rankings of the mean scores, taken separately by task and by teacher/teaching assistant. The teachers reported the force *Other Adult*, that is, the teaching assistant, to be highest of any force in amount of influence on both Task 1 and Task 2, and also the most positive in direction of influence on Task 3. The teaching assistants, for whom this force represents the teacher, returned the compliment and also scored it highest on all three tasks. This reciprocal appreciation suggests good rapport within the classroom and mutual cooperation on common goals. The *Program Advisor* was also viewed as both powerful and positive—the rank of importance scores, on both Task 1 and Task 2, was fourth for teachers, third for teaching assistants; and the degree to which this influence was judged to be positive, on Task 3, was second only to *Other Adult*, for both groups. The *Program Director* was judged eighth in extent of influence by teachers and fourth by teaching assistants, but both groups judged this force as fourth highest in the degree to which the influence was positive.

The differences in the pattern-of-importance ratings can be seen more explicitly by reference to the Task 2 mean scores in Table 1. Since the total number of points assigned to all 13 forces is 100, the number of points and the percentage of total points is the same for any one force or for the sum of any number of forces. By adding the Task 2 points for forces 11, 12, and 13, we see that the number of points assigned to all three Follow Through components was 33.2% of the total for teachers, and 49.7% for teaching assistants. It appears that the differences observed between teaching assistants in their respective patterns-of-importance rankings relate not simply to differences in the way the two groups perceive individual Follow Through components, but actually reflect a clear-cut difference in their perceptions of the program as a whole.

These differences between teachers and teaching assistants in their estimations of the influence of the Follow Through program and its components on the classroom can be understood in terms of role differences in the social-psychological impact of Follow Through. The advent of Follow Through created a new role for the teaching assistant, and she sees the inservice training and other assistance provided by program personnel as important resources in fulfilling this role in the classroom. But the teacher was there before Follow Through, and will continue to rely on/be influenced by other elements, such as the *Curriculum* and the *Social Environment*. For instance, the teacher attaches greater importance to the curriculum because she has to take greater responsibility with respect to roles and mandates built into the educational system independently of Follow Through.

Using the EFI for Force Field Analysis

Two Dimensions of the EFI

The strength of the EFI lies not so much in its ability to measure attitudes in specific areas as in its ability to reflect two different aspects of influence, power and affect, simultaneously. This feature is essential for charting a force field.

Plotting a Force Field

Force field analysis uses a plot locating each force along the two dimensions of influence, power and affect. For example, in Figure 1 there are two plots, each representing a different school district. The two axes represent the two dimensions: vertical for power, horizontal for affect.

Each district's own norms were used to calculate z-score coordinates along each dimension: Task 1 scores were used for power, Task 3 scores for affect (Task 2 scores were not utilized). To calculate the coordinates along the power dimension, all Task 1 ratings were totaled, across all raters within the district and across all 13 forces, to get an overall mean; the totals for the 13 forces were used to calculate the overall standard deviation; then, for each force separately, the mean across all raters was subtracted from the overall mean and divided by the overall standard deviation to yield a z-score deviation. The same procedure was used for calculating the affect coordinate using Task 3. These two z-score deviations were then used as the coordinates for locating each force as a point on the plot. All processing, including charting, was carried out by computer.

Forces that appear in the upper right-hand portion of the grid are those rated by teachers as having the strongest and most positive influence. Those forces located in the upper left-hand quadrant are also strong, but exert less positive influence. The EFI yields unique force field patterns for different school districts. In District C, the *Principal* and *Program Director* exert positive influence. In District E, both the *Principal* and *Program Director* are perceived as less influential.

Significance and Use of Force Field Analysis

The EFI can be used to identify patterns of forces among schools within a district. Figure 2 shows how six principals were rated by their teachers in District E. Overall, the principals in this district were rated as having a low, slightly positive influence. When individual principals are plotted, considerable variability is evident.

This procedure yields important information. A pattern of forces may be viewed within a particular school district, or specific forces may be examined across various schools and districts. Moreover, comparisons can be made on the basis of local, regional, or national norms. At the Far West Laboratory this information is being used to assist with program implementation and program improvement.

The force field pattern of an individual district may point up a problem with respect to some element in the educational setting. The influence of teachers may be perceived as low or not too positive, indicating perhaps a need for inservice training. Parents may be viewed as unimportant or a negative influence, suggesting the need for

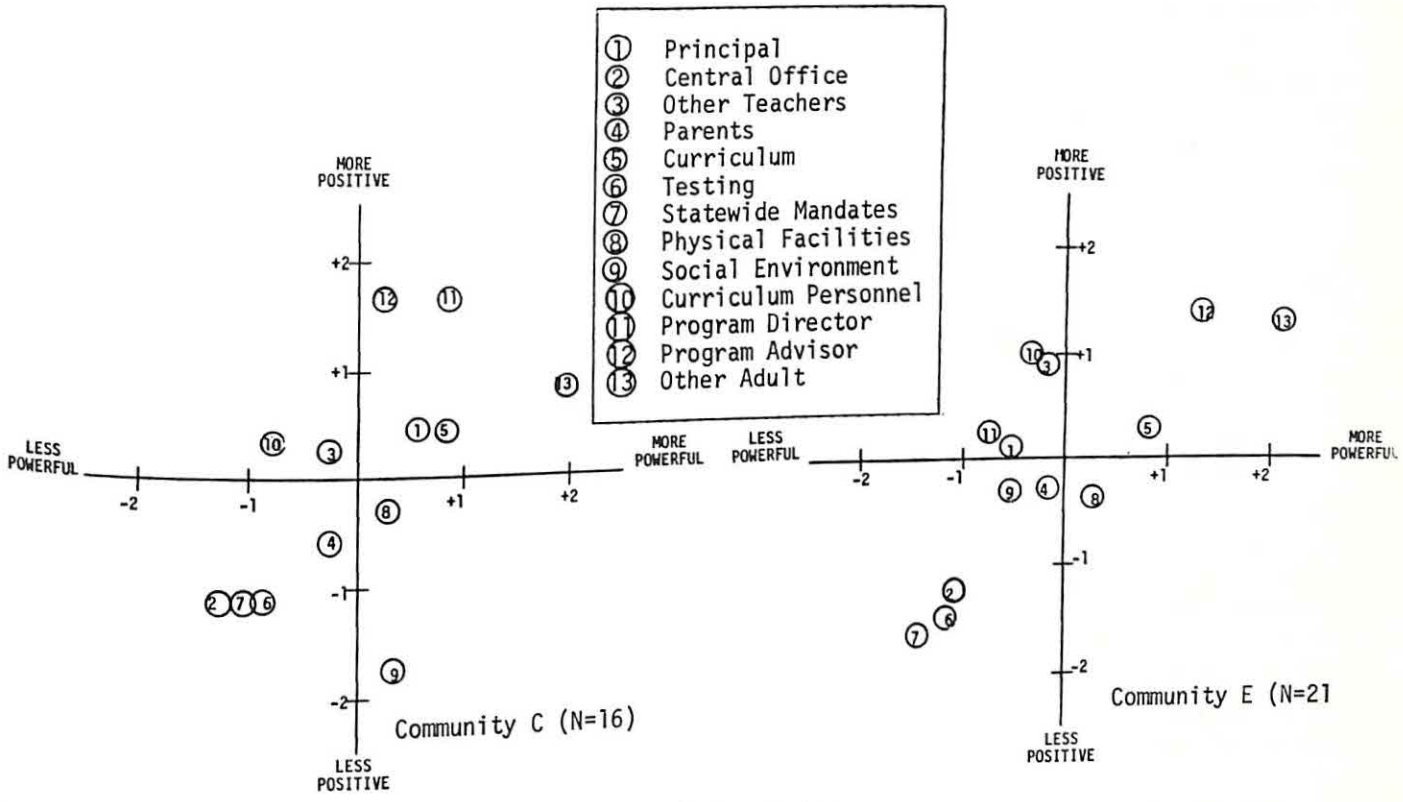


Figure 1.—Plots of *z*-Scores of Forces that Influence Teachers in Two Communities

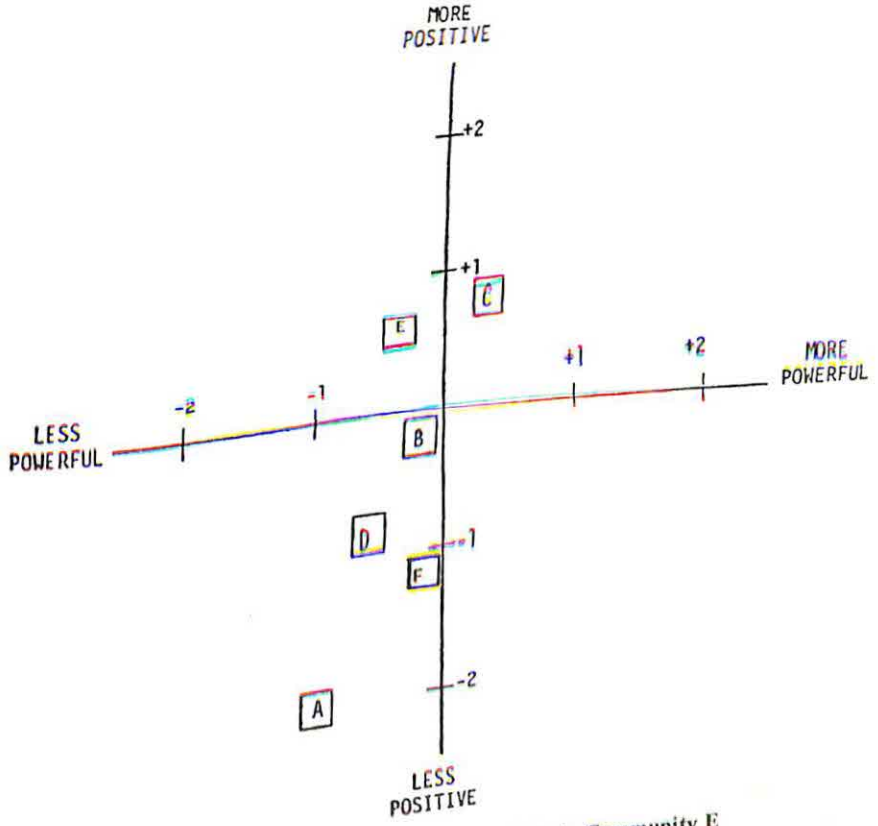


Figure 2.—Plots of *z*-Scores for Six Principals in Community E

community-oriented efforts. It may be that an important resource, such as the Program Advisor, who is responsible for inservice training, is viewed as positive but weak in influence, suggesting thte need for special training for this important staff trainer. Or it may be that a resource is identified as being strong but not very positive in influence, suggesting a need for a more basic reorientation of the direction taken by that resource.

Specific information about such conditions provided by force field analysis can be a first step in dealing with the underlying problems. For example, the information depicted in Figure 2 is a revealing commentary about the way principals in the various schools of a district are relating to the work of teachers. This information was actually used as the framework for a series of on-site workshops conducted for those principals that dealt with the principal's role and function as it relates to the educational program being implemented in the district.

Currently, the most valuable application of forces data is in educational planning, and especially with reference to preservice and inservice training of teachers.

Other important uses of the forces data allow an educational change-agent to monitor the effects of programs on the system. In such a way, district staff responsible for staff development can be monitored, with regard to teacher receptivity. Combined with other implementation data such as how much children are learning and how teachers are performing in the classroom, the EFI data can assist in the explication of the change process.

Further, as we learn more about how a packaged program is best delivered, implemented, and institutionalized within a school district, the EFI technique can be of tremendous value. It can help to document a school district's pre-existing conditions in an analytic way. Such conditions can more adequately be studied as they affect the implementation of a specific program. Ultimately patterns of conditions can be linked with delivery strategies to achieve maximum effectiveness in program implementation.

Summary and Conclusions

The EFI was constructed to reflect adequately the concerns of teachers. It provides information about the patterns of influence exerted on the teaching/learning process by significant elements in the educational setting. This information is useful in meeting the needs of teachers, in monitoring program implementation, in setting priorities, and in evaluating program impact. From the summary of data as presented below, it is evident that the EFI instrument is both practical and valid:

–The items on the instrument correspond to actual, tangible forces that can be considered and, if necessary, modified to the advantage of the teacher, the program, and, thereby, the educational process.

–Teachers and teaching assistants found the instrument simple and easy to understand.

–The scores assigned to forces were clearly related to the objective relationship of the forces to the work of the classroom: e. g., the closer the force to the classroom, the greater and more positive the influence attributed to it. The effect due to the objective relationship of a particular force to the work of the classroom was far greater than that due to the effect of professional role status, as to teacher/teaching assistant.

FOOTNOTES

1. The authors are grateful to Dr. Glendon P. Nimnicht for originally pointing out the need for this technique and for his help in conceptualizing the problem; and to Dr. Stephen Sheldon for his help in data analysis.

2. Follow Through is a federally funded program that offers comprehensive services to children from kindergarten through third grade. A sponsor, such as, in this case, the Far West Laboratory for Educational Research and Development, works with participating school districts to implement a specific instructional model, by providing curriculum materials, inservice training for local staff, etc.

REFERENCES

1. Bavelas, A., "Morale and the Training of Leaders," in G. Watson (ed.), *Civilian Morale,* Houghton Mifflin, Boston, 1942.
2. Bentley, R. R.; and Rempel, A. M., *Manual for the Purdue Teacher Opinionaire,* Purdue Research Foundation, West Lafayette, Ind., 1970.
3. Cook, W. W.; Leeds, C. H.; and Callis, R., *Minnesota Teacher Attitude Inventory,* Psychological Corporation, New York, 1951.
4. Dreeban, R., "The School as a Workplace," in R.M.W. Travers (ed.), *Second Handbook of Research on Teaching,* Rand McNally, Chicago, 1973, Ch. 14.
5. Lewin, K., *A Dynamic Theory of Personality,* McGraw-Hill, New York, 1935.
6. Lewin, K., "Forces behind Food Habits and Methods of Change," *Bulletin of the National Research Council,* 108: 35-65, 1943.
7. Lewin, K., *Resolving Social Conflicts,* Harper & Brothers, New York, 1948.
8. Lewin, K.; Lippit, R.; and White, R., "Patterns of Aggressive Behavior in Experimentally Created 'Social Climates'," *Social Psychology,* 10: 271-299, 1939.
9. Nimnicht, G. P., *The Responsive Educational Program,* Far West Laboratory for Educational Research and Development, San Francisco, 1973.
10. Rayder, N.; Ng, P.; and Rhodes, A., *Implementation of the Responsive Program: A Report on Four Planned Variation Communities,* Far West Laboratory for Educational Research and Development, San Francisco, 1973.

THE EDUCATIONAL FORCES INVENTORY: PSYCHOMETRIC PROPERTIES

NICHOLAS F. RAYDER
BART BODY
Far West Laboratory for Educational Research and Development
San Francisco, California

ABSTRACT

The Educational Forces Inventory (EFI) is a technique that charts the influence of elements in the educational setting from the point of view of the classroom teacher. It uses ratings collected individually from teaching adults, but is especially useful for mapping situational as opposed to psychological variables. Here the validity of the EFI is explicated by the findings of two studies carried out in sociological contexts: (1) for pairs of respondents, the closer their positions within the field of forces, along field-relevant dimensions of either geographical-organizational distance or work-role category, the greater the similarity in the EFI patterns generated by them; and (2) supervisory district personnel could match individual schools to teacher-generated force field patterns with a level of accuracy well above chance, and the accuracy in matching was markedly better when it was done by personnel functionally closer to the classroom. In two other studies, it was found that the size of reliability/stability correlation ratios for the patterns on EFI tasks increased with the size of the organizational unit: over a one-year interval the averages over three tasks were: .62 for individuals; .65 for classrooms; .82 for schools; and .93 for districts. These data on the psychometric properties of the EFI are interpreted as favorable indications for its use in formative evaluation of educational programs, and especially for improving the work setting of classroom teachers.

THE EDUCATIONAL FORCES Inventory (EFI) charts the elements or forces in the educational setting from the point of view of the classroom teacher. Forces that represent salient aspects of the given educational setting are assessed by teachers in terms of their relative influence upon classroom process along each of two dimensions: power, or the amount of influence exerted; and affect, or the extent to which influence is felt to be positive or negative. These ratings are then used to generate coordinates for mapping each force over a two-dimensional field. This representation of the social-psychological context of teachers provides valuable data on specific elements of the educational setting.

A previous paper by Rayder and Body (2:26-34) described the rationale, initial development, and field-testing of the EFI, and also reported on its concurrent validity with reference to the Purdue Teacher Opinionaire, an older, established instrument with an essentially psychological, intra-individual focus. This paper reports new data that has bearing on the inter-individual or sociological ramifications of the EFI. Specifically, it will discuss: (a) the patterns of agreement among individuals who vary along important dimensions of social relationships; and (b) the degree of correspondence between EFI force fields and the independent assessments of knowledgeable observers. In addition, extensive data on the reliability/stability characteristics of the EFI will be presented.

Description of the EFI[1]

Elements of the EFI

The EFI is based on forces in the teacher's social-psychological field which have a significant influence on morale and classroom effectiveness. There are 13 forces, selected and delineated to correspond to salient elements of the classroom teacher's work-space as they overtly and tangibly present themselves to the teacher. Ten of the forces are relevant to public schools generally:

1. *Principal* of the school
2. *Central Office* administrative personnel
3. *Other Teachers* in the school
4. *Parents* of children in the class
5. *Curriculum* prescribed by the district
6. *Testing* programs
7. *Board of Education*
8. *Physical Facilities* available
9. *Social Environment* of the community
10. *You, Yourself*

The three other forces refer to salient features of the Responsive Education Model Follow Through Program:[2]

11. *Program Director*—coordinator of the Follow Through Program within the district: responsible for administration, community organization and policy matters.

12. *Program Advisor*—delivers the program to the classroom with inservice training and in-class assistance: each advisor is responsible for about ten classrooms.
13. *Other Adult* in the classroom—teacher or teaching assistant: a teaching assistant in this model is a full-time, paid paraprofessional assigned to the classroom and engaged in teaching activities.

Recording Teachers' Attitudes on the EFI

In completing the instrument, the respondent—teacher or teaching assistant—is asked to evaluate the set of 13 forces by carrying out three successive tasks:

Task A: Each force is rated on its importance in influencing teaching, on a scale of 0 - 9. A rating of 0 indicates no influence, a rating of 9 indicates a strong influence of either positive or negative effect.

Task B: Each force is assigned a weight according to its relative importance in influencing teaching. A total of 100 points are distributed among the 13 forces, in direct proportion to their amount of influence. Any pattern of assignments is permissible; for instance, the respondent may choose to distribute the points evenly among the 13 forces or allocate all 100 points to just one or two of them.

Task C: Each force is rated on its positive/negative effect on teaching, on a scale of 1 to 9, with a rating of 1 indicating strong negative influence, and a rating of 9 indicating strong positive influence.

Plotting a Force Field Pattern for Schools

Force field analysis generates a plot locating each force along the two dimensions of influence, power and affect. For example, Figure 1 shows two plots, representing the EFI patterns constructed from the aggregate of teachers in each of two different schools in the same district. The two dimensions of the page represent the two dimensions of influence assessed: vertical for power, horizontal for affect. Task A scores were used for power, Task C scores for affect. Scores of Task B overlapped considerably with those of Task A and were therefore not utilized.

In each plot, the position of each force is determined by a pair of coordinate values that correspond to *z*-score deviations of the school relative to the whole district. For instance, to calculate the Task A (power) coordinate of *Principal*, the mean rating for *Principal* on Task A, over all respondents in the schools, is subtracted from the mean rating for *Principal* over all respondents in the district. This difference is then divided by the standard deviation of the set of means for *Principal* for all schools in the district. The procedure for Task C (affect) is analogous.

Forces that appear in the upper right-hand portion of the grid are those rated by teachers as having the highest and most positive influence on their teaching in the classroom. Those forces located in the upper left-hand quadrant are also rated as having relatively high but less positive influence. Two patterns of influence are evident for the two schools. In School A the *Principal* exerts a strong positive influence; in School B the *Principal* is less influential. In both schools, the *Teaching Assistant* is seen as having a strong influence; in School A, however, this influence is perceived as distinctly less positive than it is in School B.

Practicality, Validity, and Reliability of the EFI

Preliminary Indications

The initial paper on the EFI reported data on its practicality and validity:

—Teachers understood and accepted the task of completing the instrument, as evidenced by their ability and willingness to respond. [See (2:29).]

—Ratings assigned were primarily related to the external factors in the school/work settings, as opposed to individual characteristics of the respondents, such as professional role. [See (2:31).]

—Concurrent validity was demonstrated with reference to the Purdue Teacher Opinionaire, an older, established instrument that was administered concurrently. [See (2:31).]

More Data on Validity and Reliability: Four Studies

The four studies reported here, with new data in new contexts, enlarge upon these preliminary indications of validity and also present data on reliability/stability. Study I and Study II deal with aspects of validity in the interpersonal context, using data obtained in a program-wide survey of all teachers and teaching assistants working with the Responsive Educational Program in the spring of 1973. The other two studies deal with aspects of the reliability/stability of force field patterns aggregated over individuals in organizational units of various levels: Study III uses data obtained from a single district, on two occasions four weeks apart, in early 1974; Study IV is based on data collected in program-wide surveys in spring 1973 and spring 1974.

Validity of EFI Patterns Tested in Two Interpersonal Contexts

Study I: Validity with Reference to Patterns of Agreement among Respondents

It follows from the work of Lewin (1) that the closer to each other two observers are located within a field of social-psychological forces, the more alike they will be in the way they perceive these forces impinging on them and on their work. For classroom teachers, two important dimensions may be used to define relative position in the field of forces corresponding to the educational setting: (1) organizational operating units such as class, school, and district; and (2) work roles such as teacher and teaching assistant. Since perception of the field of forces is dependent upon the observer's position within it, one way to check whether the EFI technique generates valid results is to examine the

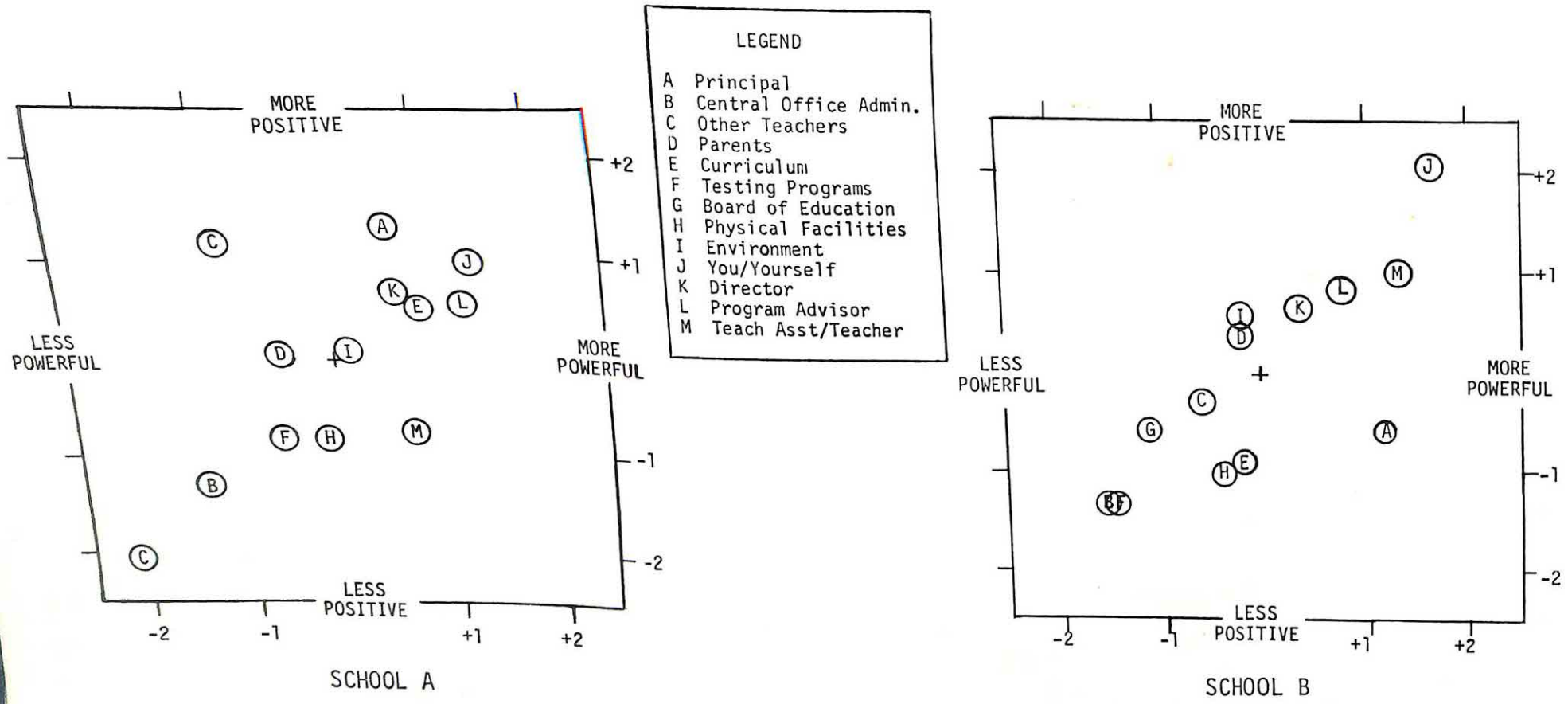


Figure 1.—Plots of z-scores of Forces that Influence Teachers in Two Schools within the Same District

patterning of data along these two dimensions. If the patterns accurately reflect reality, it is to be expected that the nearer any two respondents are to each other along these dimensions, the more alike they will be in their responses such about the field. In other words, a validity criterion is such that, with respect to organizational units, it will be the case that, on the average:

—pairs of respondents teaching in the same classroom will be more alike in their responses to each other than pairs teaching in different classrooms;
—pairs within the same schools will be more alike than pairs in different schools;
—pairs within the same school district will be more alike than pairs in different districts.

Similarly, with respect to professional roles, it will be the case that, on the average, pairs of either teachers or of teaching assistants will be more alike in their responses than mixed pairs of teachers/teaching assistants.

The Responsive Education Program provides systematic variations along both the dimension of organizational operating unit and the dimension of professional role. The program extends over 14 school districts in 12 states, grades kindergarten through third, thus providing an opportunity to compare individuals that are both administratively and geographically distant. At the same time, since one teacher and one teaching assistant are assigned to each classroom, it is possible to compare individuals who work quite closely together. Moreover the pattern of parallel assignments provides the opportunity to compare easily and meaningfully across the dimension of professional role as exemplified by these two categories, teacher and teaching assistant.

An EFI survey was directed to all of the roughly 700 teachers and teaching assistants working within the Responsive Education Program during the school year 1972-73. Of the 604 returns, 29 (mostly from teaching assistants) could not be fully processed: 24 because they could not be identified as to the respondent's classroom and/or work role; and another 5 because they were incomplete on more than one task. There remained 515 valid returns, 304 from teachers, and 211 from teaching assistants.

For each of the three tasks the average correlation was computed for teacher/teaching assistant pairs: (1) within the same classroom; (2) in different classrooms within the same school; (3) in different schools within the same district; and (4) from different districts. Average correlations were also computed separately for pairs of teachers and for pairs of teaching assistants for the last three categories. Since correlation coefficients are not additive, the average correlation for a particular category of pairs was computed indirectly as follows: the product-moment correlation coefficients were computed for all pairs within the category, then converted to their z-score equivalents; next, the mean of the z-score equivalents was computed and converted back to a correlation coefficient. Because of the excessive number of potential pairings in the "Different Districts" category (around 40,000), the average correlations here were based on a sample of 1,000, randomly selected out of all possible pairs. Where a task was not completed, the respondent was dropped from the pairings for that task but included in those tasks that were complete. All processing, including matching, calculating, and randomizing, was carried out by computer. The results are presented in Table 1, and charted in Figure 2.

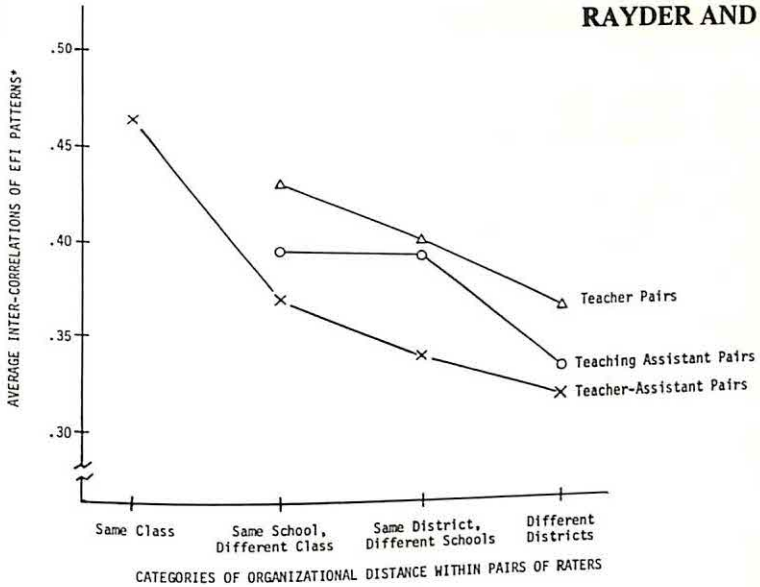
The patterns of responses on the EFI reflected a direct relationship between relative closeness of the organizational operating unit of class, school, and district on the one hand, and congruity of perceptions on the other. This trend prevailed throughout and was most evident for the Teacher/Teaching Assistant pairs (different-role) on Task A, where the correlations diminished in regular steps from .49 to .31. The differences were less dramatic for either Teacher/Teacher or Teaching Assistant/Teaching Assistant (same-role) pairs mainly because one of the end points on the unit's dimension—the "Same Class" category, denoting two respondents teaching in the same classroom—was missing, thus reducing the range of variation.

From these data it can also be said that the perceptions of the field of forces are more similar within each of the two job categories than across them: correlations for the different-role pairs are generally lower than corresponding ones for same-role pairs. In Figure 2, this relationship is reflected in the fact that the line representing different-role pairs is the lowest of the three.

With respect to both these dimensions, geographical-organizational distance and professional role, the EFI reflects real differences in the social-psychological field in consistent ways. For pairs of respondents, the closer their positions within the field of forces, along either dimension, the greater is the similarity in their patterns of assessing the set of forces on the EFI.

Study II: Validity of the Force Field Pattern: Recognition

To test whether the EFI leads to valid descriptions of the constellation of forces in a particular school, force field plots may be compared to independent assessments made by knowledgeable individuals. Such a comparison was carried out using a forced-choice procedure. A force field plot was constructed for each of the 61 schools in the 13 districts in which the Responsive Education Program functioned in at least two schools (in the 14th district all Follow Through classrooms were in a single school). These plots, identified only by a code number, were then distributed to the respective districts, and the district program personnel involved in supervisory-level positions (program directors and program advisors) were asked to identify the



*Correlations averaged across respondents within each category of organizational distance and across three tasks—see also text and Table 1, far-right column.

Figure 2.—Relationship between Organizational Distance within Pairs and Similarity of EFI Patterns for Average of Three Tasks in Pairs of Teachers, Teaching Assistants, and of Teacher-Teaching Assistant

Table1.—Average Correlations between EFI Patterns of Pairs of Respondents, by Respondents' Professional Roles, and by Category of Organizational Distance between Roles

Category of Organizational Distance Within Pairs	Avg. Correlations Between EFI Patterns of Pairs of Respondents				
	Minimum Number of Pairs	Task A	Task B	Task C	Average of Three Correlations
	Same-Role Pairs -- Teachers				
Same School, Different Classes	712	.44	.48	.37	.43
Same District, Different Schools	2211	.42	.46	.29	.40
Different Districts	1000	.36	.43	.30	.37
	Same-Role Pairs -- Teaching Assistants				
Same School, Different Classes	376	.34	.44	.37	.40
Same District, Different Schools	1026	.35	.42	.38	.39
Different Districts	1000	.32	.38	.31	.33
	Different-Role Pairs -- Teacher-Teaching Assistant				
Same Class	139	.49	.47	.44	.47
Same School, Different Classes	1119	.36	.40	.34	.37
Same District, Different Schools	2912	.33	.36	.30	.33
Different Districts	1000	.31	.34	.31	.32

school represented by each plot. The results are presented in Table 2.

The six schools in District 4 had been divided into three sets of two, in correspondence with the program advisor assignments. By chance alone, the expected number of correct matches, or "hits" would have been one for each set of schools to be matched, or 15 in total. The actual number of hits was 27 out of 61 possible schools. Thus, the improvement over chance is significant, both practically and statistically ($z = 2.3, p < .02$).

Some districts had greater accuracy in matching than others. In every case where there were only two schools to choose from, the choices were correct. This was the case with Districts 1, 2, and 3, in each of which only two schools were involved, and in District 4, where the set of six schools had been divided into three sets of two, and the matching for each pair was done by the program advisor working with the particular pair. In contrast, the number of hits scored in the four districts with six or more schools to choose from was only slightly better than expected by chance alone: expected hits, 4 of 27; actual hits, 6 of 27. Where the number of schools to choose from was four or five, the improvement over chance was intermediate between these two: expected hits, 5 of 22; actual hits, 9 of 22.

In general, the matching was done by the program director, whose district-wide responsibilities were at least one step removed from the classroom. In the smaller districts, however, the program director might also double as program advisor with in-class activities. And in District 4, the matching was done by the particular program advisor involved in each subset of two schools. Evidently, accuracy in matching was directly related to proximity to the classroom in terms of administrative distance and work role.

To explore this notion further, another round of judgments was collected. The schools in three of the larger districts (8, 12, and 13) were partitioned by the program advisor in charge, who was requested to carry out the matching (Districts 10 and 11 could not be polled in this round because of particular difficulties existing there at the time, such as excessive turnover of staff; District 9, though it had Follow Through in 5 different schools, was actually one of the smaller districts, with only 14 classrooms and 2 program advisors; and District 12 could not be included because each of the program advisors was involved in only one school). The results are presented in Table 3.

By chance alone, the expected number of hits would have been 8 of 17, and the actual number of hits was 15. Table 4 summarizes the data from both rounds of matching in terms of the number of schools in the set to be discriminated.

There were 29 cases where matching involved selecting from a set of two or three schools at a time: 12 in the first round, in Districts 1, 2, 3, and 4; and all 17 in the second round. For these 29 the number of hits expected from chance alone was 14, and the actual number scored was 27. Thus, the index of improvement over chance, Kappa = .87 for sets of 2-3 schools; for sets of 4-5 schools, Kappa = .22; and for sets of 6-9 schools, Kappa = .09. Clearly, the discriminations were much better for smaller sets, where the rater was operationally closer to the scene.

Table 2.—Number of "Hits" in First Round of Matching Schools to Force Field Plots, by District

District	Number of Schools	Number of Expected "Hits" by Chance Alone	Number of Actual "Hits"	Number of Actual "Hits" in Excess of Chance
1	2	1	2	1
2	2	1	2	1
3	2	1	2	1
4*	6	3	6	3
5	4	1	1	0
6	4	1	2	1
7	4	1	2	1
8	5	1	1	0
9	5	1	3	2
10	6	1	1	0
11	6	1	0	-1
12	6	1	3	2
13	9	1	2	1
—	—	—	—	—
Total 13	61	15	27	12

Instead of the six plots being sent to the district's program director, they were sent, two each, to the three program advisors who had been working directly with the two schools for a year or more.

Reliability/Stability of EFI Patterns of Operational Units at Different Levels

Reliability vs. Stability

An important aspect of instruments such as standardized achievement tests or personality inventories is their test-retest reliability, meaning the extent to which an individual's performance is constant over a period of time that is long enough to mitigate the effects of memory, practice, momentary set, etc., but not so long that the abilities or the personality of the individual have changed appreciably. The social-psychological field depicted by the EFI is constantly changing because of changes in the external environment as well as in the individual raters. But if the force field plots are based on a larger number of raters, the idiosyncracies of individuals tend to average out, leaving the effects of external change to be reflected all the more clearly. Therefore, in referring to consistency of EFI patterns from one testing occasion to another, it is more appropriate to think in terms of "stability" rather than "reliability."

Study III: Stability of Individual and Group Patterns Over Four Weeks

Arrangements were made for two administrations of the EFI in one of the districts. All teachers and teaching assistants in District 4 were asked to complete the EFI, in

Table 3.–Number of "Hits" in Second Round of Matching Schools to Force Field Plots, by Program Advisor within District

District Number	Program Advisor Number	Number of Schools/ Plots to be Matched	Number of Expected "Hits" by Chance Alone	Number of Actual "Hits"	Number of Actual "Hits" in Excess of Chance
8	1	2	1	2	1
	2	2	1	2	1
	3	2	1	0	-1
12	1	3	1	3	2
	2	2	1	2	1
13	1	2	1	2	1
	2	2	1	2	1
	3	2	1	2	1
Total (3)	(8)	17	8	15	7

Table 4.–Number of "Hits" in Round One and Round Two Combined, by the Number in the Set to be Matched

Number in the Set to be Matched	Number of Schools to be Matched	Number of "Hits" Expected by Chance Alone	Number of Actual "Hits"	Number of "Hits" in Excess of Chance	Kappa*
2 - 3	29	14	27	13	.87
4 - 5	22	5	9	4	.22
6 - 9	27	4	6	2	.09

* Kappa = Hits in Excess of Chance ÷ Potential Improvement.

Table 5.–Stability of Individual Patterns of Responses over Four Weeks, by Task

	Task A	Task B	Task C	Average Over Three Tasks
Number of Correlations Averaged	37	27	29	93
Mean of Corresponding z-Scores	.68	.88	.84	.78
Corresponding Correlation Ratios	.59	.71	.68	.66

the winter of 1974. No mention of a retest was made. After exactly four weeks, the procedure was repeated.

Of the 58 teachers and teaching assistants in the district, 42 successfully responded to the test and retest. Product-moment correlations were computed between the two sets of responses obtained from each individual on each of the three tasks. The average correlation ratio for each task was computed over the set of all teachers. This procedure involved: (1) calculating correlation ratios for each individual; (2) converting correlation ratios to z-score equivalents; (3) taking the mean; and (4) converting back to correlation ratios. The results are presented in Table 5. The average stability coefficients ranged from .59 to .71.

An index of test-retest stability for the district as a whole was also computed for each of the three tasks. Item responses were averaged across individuals, and the correlation ratios between the two group patterns thus obtained on the two occasions were: Task A, .95; Task B, .98; and Task C, .98.

Study IV: Stability of Individual and Group Patterns Over a Year

In the usual psychological test the responses on each item are added or averaged to get a score for each individual: with the EFI the responses of all individuals are averaged to get a pattern for the unit. In either case, the more elements that are averaged, the higher the reliability. In the EFI, the number of individuals averaged to get a particular pattern is limited only by the number of respondents in the unit of interest.

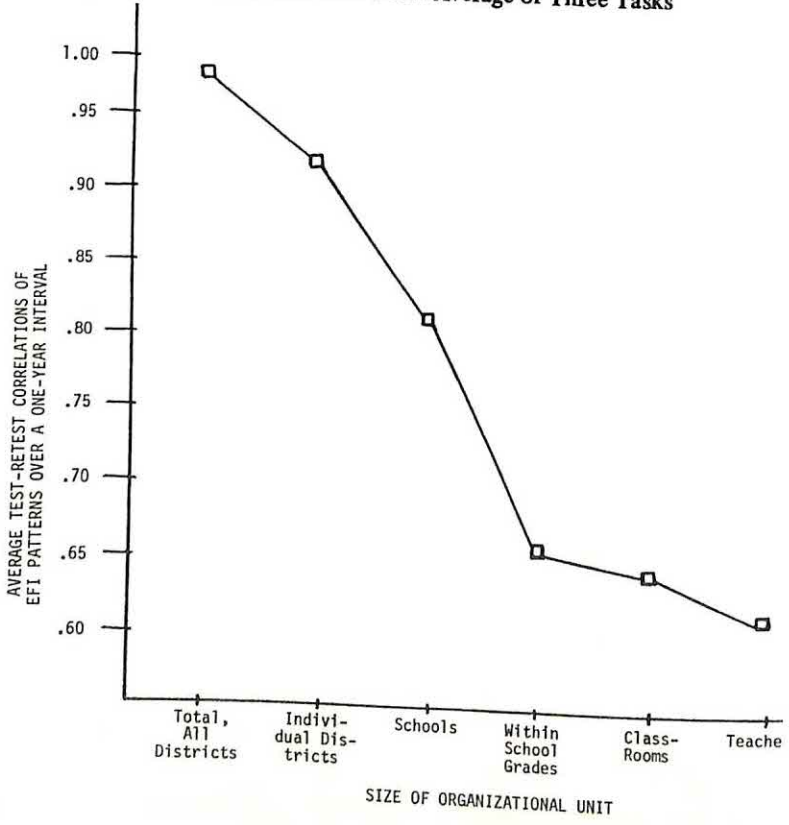
The average test-retest correlations obtained for the two classes, individuals and the district as a whole, represent two points along a continuum of unit size that includes: individuals, classes, schools, districts, geographic areas, etc.

Table 6.—Average Test-Retest Correlations Over a One-Year Interval for EFI Tasks A, B, and C, Aggregated by Various Levels of Organizational Unit

	Task A	Task B	Task C	Average Over Three Tasks
Total, All 13 Districts				
No. of Elements Averaged	1	1	1	3
Mean of z-Scores	2.75	3.05	2.86	2.89
Corresponding Correlation Ratios	.99	1.00	.99	.99
Individual Districts				
No. of Elements Averaged	13	13	13	39
Mean of z-Scores	1.66	1.69	1.73	1.69
Corresponding Correlation Ratios	.93	.93	.94	.93
Standard Deviation of z-Scores	.26	.33	.20	.27
Schools				
No. of Elements Averaged	52	52	52	156
Mean of z-Scores	1.14	1.20	1.09	1.14
Corresponding Correlation Ratios	.81	.83	.80	.82
Standard Deviation of z-Scores	.33	.51	.42	.43
Grade Levels Within Schools				
No. of Elements Averaged	131	129	131	391
Mean of z-Scores	.78	.89	.73	.80
Corresponding Correlation Ratios	.65	.71	.62	.66
Standard Deviation of z-Scores	.39	.51	.41	.45
Classrooms				
No. of Elements Averaged	119	119	119	357
Mean of z-Scores	.74	.84	.72	.77
Corresponding Correlation Ratios	.63	.69	.62	.65
Standard Deviation of z-Scores	.37	.52	.39	.44
Teachers				
No. of Elements Averaged *	118	118	118	354
Mean of z-Scores	.74	.78	.67	.73
Corresponding Correlation Ratios	.63	.65	.58	.62
Standard Deviation of z-Scores	.33	.56	.40	.44

*There were 119 teachers, but each of the three tasks was left incomplete by one teacher, a different one on each task.

Figure 3.—Relationship between Size of Organizational Unit and Stability of EFI Patterns Over One Year: Average of Three Tasks



The figures for these two points are consistent with the notion that test-retest stability of the EFI increases with increase in the size of the unit surveyed.

To test the generality of this effect, a special analysis was carried out for program-wide data collected in two successive years. Indices of year-to-year reliablity/stability for several points along the continuum of unit size were calculated. The data were collected in surveys conducted in the spring of 1973 and of 1974 as part of the regular monitoring of program implementation. In both years, the survey was directed to the 300 teachers and the 300 teaching assistants in the Responsive Education Program. However, District 4 was not included in the spring 1974 survey because it had already been surveyed twice a few months earlier as part of Study III. These two administrations of the inventory yielded year-to-year data points for 13 school districts, with 52 schools, with 132 grade levels within schools, and with 119 classrooms.

In a strict sense, no classroom is the same from one year to the next. For purposes of this analysis, however, a classroom was considered the same if the school, the grade level, and the teacher wre the same. There were 119 such classrooms for which the returns of at least the teacher could be referred to the same individual at both time points. For 45 of these, the teaching assistant in the classroom was the same and also returned an EFI form at both time points. Attrition was due to respondents opting to remain anonymous or to their decision not to participate. These factors reduced the number of elements that could be included in the analysis from among the individual and the classroom categories. In the case of the larger units, where aggregation was both feasible and meaningful without identification of individual respondents, year-to-year comparisons could be made without regard to whether any or all of the respondents were the same at the two time points.

Test-retest correlations were computed for each task: for the responses aggregated over all 13 districts taken together, and individually for each of the 13 districts, 52 schools, 132 grade levels within schools, 119 classrooms teachers and current teaching assistant where available) and 119 teachers. The data are presented in Table 6.

There is a consistent decrease in the size of the correlations as we move from program to district, to school, to grade level, to classroom, to teacher. The size of the difference varies, but the direction of difference is always the same: the smaller the unit, the lower the average correlation. This trend is charted in Figure 3.

It is of special interest that the stability indices are lower for teachers than for classrooms. If the EFI reflected only personal, rather than inter-individual factors, we would expect just the reverse. The sample of teachers was defined so as to insure that the scores at both time points would reflect the same individuals, whereas the scores for classrooms represent averages of teachers and teaching assistants, and the latter are for the most part different individuals at the two time points. This result is another indication that the EFI faithfully reflects the reality and the constancy in the educational setting, rather than any purely personal or a-situational factors impinging on the respondents.

Summary and Conclusions

In light of these four studies, the psychometric properties of the EFI instrument are seen to be very much in keeping with the purpose for which it was originally designed: systematically mapping the educational setting considered as the work-space or the psycho-social field of forces of the classroom teacher.

Study I, and in some sense *Study II* as well, demonstrated that the degree to which the patterns of two different raters are similar to each other is directly and consistently related to important dimensions which define the environmental setting considered as the work-space of the classroom teacher—the dimensions of administrative-geographic distance and professional role. In other words, the EFI does differentiate the field of forces within the teacher's work-space along lines that are significant to the teacher.

Study II showed that force field patterns generated by the EFI can be referred to the educational setting—that these patterns are confirmed by the independent perception of knowledgeable individuals.

Study III and Study IV showed that the EFI patterns are highly stable for schools and districts (*r*'s of .80-.94) and moderately high for classrooms and grade levels (*r*'s of .62-.71), and that the reliability increases with increase in the size of the unit sampled.

In education, as in any institutional or group enterprise, individual and group change is most effective when norms or standards regulating behavior are changed. When a norm is changed, group members change their behavior to conform to the new norm. On the other hand, attempts to change group or organizational behavior by changing individual behavior often results in resistance to change, particularly when an individual perceives that the change is not endorsed by his peers. Thus, the primary focus in the development of educational programs and the improvement of the educational setting should be normative change, with individual change a by-product.

In and of itself, the EFI survey does not bring about changes in the educational setting, unless we count the changes that come with an opportunity for harried teachers to ventilate. Once the situation has been specified in terms of an EFI pattern, however, remedial action regarding factors related to norms and/or the work setting may be obvious; at least the options will have been more clearly defined.

At the Far West Laboratory, force field patterns are being used as an important component of the process of program evaluation. Unlike performance-based criteria such as child test scores or teacher behavior indices, they relate directly to what help is being or can be provided the educational practitioner—teacher, teaching assistant, prin-

cipal, curriculum specialist, program developer, or other change-agent. A report elaborating the specifics of using the EFI for formative evaluation is in progress.

FOOTNOTES

1. The description of the EFI given here reflects the format of the instrument in its current form, which was in use when the data reported here were generated. The previous publication (2:27) was based on an earlier version that differed as follows:
   a. Force #7 was *Statewide Mandates* on certifications, curriculum, grading, etc.
   b. Force #10 was *Curriculum Personnel,* such as reading specialist, art teacher, etc.
   c. Tasks were One, Two and Three, rather than A, B, and C.
   d. The first task was a ranking, rather than the Task A rating from 0 - 9.
   e. The range of ratings on the third task was 1 - 5 rather than 0 - 9.
2. For a description of Follow Through, see (2:34).

REFERENCES

1. Lewin, K., *Resolving Social Conflicts,* Harper & Brothers, New York, 1948.
2. Rayder, N. F.; and Bödy, B., "The Educational Forces Inventory: A New Technique for Measuring Influences on the Classroom," *Journal of Experimental Education,* 44:26-34, Winter 1975.

THE EFFECT OF ENCODING AND AN EXTERNAL MEMORY DEVICE ON NOTE TAKING

LINDA ANNIS
Ball State University

J. KENT DAVIS
Purdue University

ABSTRACT

College students were randomly assigned to seven note-taking and review conditions in order to determine the relative importance of the functions of encoding and either an externally provided or a personally produced memory device. Results of the post-test showed that a combination of encoding and reviewing either one's own notes or an outline of the lecture produced the best recall scores, while either personally encoding notes or being provided with a lecture outline during the lecture accompanied by "mental" review produced the least recall. The findings are discussed in terms of practical suggestions for professors and their students.

A UNIVERSAL CHARACTERISTIC of college students is that they report for class carrying a notebook in which to take notes on material presented during class lectures. Students put a great deal of effort and faith into the taking of notes presumably so they can be used later while reviewing for exams. Despite this widespread practice of taking notes, little experimental evidence exists as to the exact functions of note taking. Note taking appears to serve either or both of two functions: an encoding function in which the material heard in a lecture is transformed into a personally meaningful form, and an external memory function which serves for later review. DiVesta and Gray (1) and Howe (3) found that the encoding function was the more important of the two functions. These authors argued that too much reliance on notes as an external memory device can result in inefficient learning if the crucial encoding stage is bypassed. Howe (3) suggested that if the only function of notes were the external record of information provided by the professor during class, it would be more efficient to hand out mimeographed outlines of the lecture before class so students would be free to react to other things. Fisher and Harris (2), however, found that of the two possible functions of taking notes, the external memory device had the greatest facilitating effect on recall. They found that *Ss* who only made use of notes as an external memory device performed better on a test of recall than *Ss* who benefited from both of the functions of note taking.

The purpose of the present study was to investigate the relative importance of the note-taking functions of encoding and an external memory device, as well as various combinations of these two functions, by manipulating the note-taking and review conditions of the study. It also attempted to assess whether the better external memory de-

vice was one personally produced by the students or externally provided by the professor.

Method

Subjects

Eighty-five students enrolled in four sections of a sophomore Human Growth and Development course served as *Ss*. One randomly selected section served as a control group in which the students received no instructions regarding note taking and were provided no opportunity to review their notes prior to an examination on the content of a lecture. A 40-minute lecture on the principles of behavior modification was presented to each of the four sections by the senior author, who was also the regular instructor. Behavior modification is a topic ordinarily covered in the course in a lecture of about this length, but it was not discussed at all in any of the required readings.

Procedures

Two note-taking conditions were employed: in one condition students were instructed to take notes, and in another condition students were provided a copy of the lecturer's notes. These two note-taking conditions were paired with four different review conditions. Some students reviewed their own personally encoded notes (RON); others reviewed a copy of the lecturer's notes (RLN) which served as an external memory device; others were provided their own notes and a copy of the lecturer's notes (RON + RLN); and still others were instructed to review the material mentally (MR).

The note-taking conditions and the review conditions were combined to form the following seven treatment conditions: (1) *Ss* took notes and reviewed their own notes (N-RON); (2) *Ss* took notes and reviewed a copy of the lecturer's notes (N-RLN); (3) *Ss* took notes and reviewed mentally (N-MR); (4) *Ss* took notes and reviewed both their own notes and the lecturer's notes (N-RON + RLN); (5) *Ss* were provided a copy of the lecturer's notes and reviewed a copy of the lecturer's notes (LN-RLN); (6) *Ss* were provided a copy of the lecturer's notes and reviewed mentally (LN-MR); and (7) the control condition in which *Ss* received no instructions on note taking and were provided no review time (NIT-NRT).

Since four sections were available, an attempt was made to avoid reactivity of experimental arrangements by assigning similar treatment conditions to the same section. Classes were randomly assigned to combinations of the two most similar treatments. In class one, students were randomly assigned to conditions N-RON or N-RLN; in class two, students were assigned to conditions N-MR or LN-MR; in class three, students were assigned to conditions LN-RLN or N-RON + RLN; and the fourth class comprised the NIT-NRT control group.

On the day of the behavior modification lecture students in three of the sections were provided a packet which contained instructions and materials appropriate to their treatment condition. Students in the fourth section were not provided any instructions or materials. Following the lecture all students were asked to turn in their notes so that the instructor might assess the effectiveness of the lecture. Two weeks following the lecture a review session and an examination were administered. Prior to a 10-minute review period each student received a packet compiled according to his treatment condition. Students in the RON review condition received their own notes for review, students in the RLN condition received only a copy of the lecturer's notes to review, and students in the RON + RLN condition received both their own notes as well as a copy of the lecturer's notes. Students in the MR condition were instructed to review the lecture on behavior modification "mentally" (i. e., sit and think about the material). The control class received no time for review. All *Ss* were given a 21-point examination consisting of 10 objective questions and 5 short answer questions worth a total of 11 points.

Table 1.—Analysis of Variance for Short Answer Items

Source of Variation	SS	df	MS	F
Between Groups	135.87	6	22.65	7.7*
Within Groups	229.35	78	2.94	
Total	365.22	84		

*$F_{.01(6/60)}=3.12$

Table 2.—Analysis of Variance for Objective Items

Source of Variation	SS	df	MS	F
Between Groups	31.28	6	5.21	1.83
Within Groups	221.71	78	2.84	
Total	252.99	84		

$F_{.05(6/60)}=2.25$

Table 3.—Analysis of Variance for Total Scores

Source of Variation	SS	df	MS	F
Between Groups	164.67	6	27.44	3.84*
Within Groups	556.98	78	7.14	
Total	721.65	84		

*$F_{.01(6/60)}=3.12$

Results

A single-factor, unweighted-means analysis of variance was performed on the dependent variables of total test score (number of correct responses), number of correct responses on the short answer part of the examination, and number of correct responses on the objective items. A significant treatment effect was found for the total score ($F = 3.84$; $df = 6/78$; $p < .01$), and for the number correct on the short answer items ($F = 7.7$; $df = 6/78$; $p < .01$), but not for the number correct on the objective items ($F = 1.83$; $df = 6/78$; $p > .05$). See Tables 1, 2, and 3.

Table 4.—Means and Standard Deviations for Short Answer, Objective, and Total Scores by Treatment Condition

Treatment Condition	N	Short Answer $\bar{X}$	SD	Objective $\bar{X}$	SD	Total $\bar{X}$	SD
N-MR	10	2.8	1.14	7.3	1.83	10.1	2.38
LN-MR	10	3.0	1.25	7.7	1.49	10.7	1.89
NIT-NRT	24	3.33	1.55	6.38	1.44	9.71	2.27
N-RON	13	6.31	1.89	6.46	2.44	12.77	3.75
N-RLN	12	5.33	1.78	7.17	1.53	12.5	2.11
LN-RLN	9	4.78	2.49	7.22	.97	12	3.16
N-RON+RLN	7	5.29	1.89	8.29	1.8	13.57	3.1

The mean scores and standard deviations for the short answer items, objective items, and total scores within each of the seven treatment conditions are presented in Table 4.

Discussion

The results of this study suggest that both note-taking functions are important, but that the more important function for success in later recall is the encoding function. Examination of the three top conditions for both significant dependent variables (N-RON + RLN, N-RON, N-RLN) indicated that the most important function for success on an examination was the encoding of a personal set of notes, but that it made little difference whether the external memory device was externally provided or personally produced. Apparently the substitution of a copy of the lecturer's notes for the *S*'s own notes in the review process does not cause any interference. The fourth performing group for both dependent variables was the LN-RLN condition, which did not benefit from the encoding function but was provided with an external memory aid. The three groups with the lowest total score and the poorest recall on the short answer items either had no external memory aid or were in the control condition. Mental review does not appear to be very successful regardless of the circumstances.

Several practical suggestions for both professors and their students arise from this study. The process of personally encoding the lecture through the taking of notes is very important for success on tests of recall. Although, as this study suggests, the student's grade on an examination may depend on his skill as a note taker, very little attention has been given to training in this skill. Additional research is needed comparing various specific methods of taking notes (i. e., record main points only, record as many details of the lecture as possible) so that students can be instructed as to the most efficient and effective method for taking notes. An external memory device for review also is important, but it does not seem to matter whether this device is the same set of notes personally prepared by the student or a copy of the lecturer's notes.

REFERENCES

1. DiVesta, Francis J.; and Gray, G. Susan, "Listening and Note Taking," *Journal of Educational Psychology*, 14:8-14, February 1972.
2. Fisher, Judith L.; and Harris, Mary B., "Effect of Note Taking and Review on Recall," *Journal of Educational Psychology*, 65:321-325, March 1973.
3. Howe, Michael J. A., "Note-taking Strategy, Review, and Long-term Retention of Verbal Information," *Journal of Educational Research*, 63:285, February 1970.

ACHIEVEMENT MOTIVATION TRAINING FOR LOW-ACHIEVING EIGHTH AND TENTH GRADE BOYS[1]

KELVIN RYALS
University of Houston

ABSTRACT

An achievement motivation training course was given to 24 teachers in San Mateo County, California. They then trained 136 eighth and tenth grade students (four weekends during the fall school semester). Student training was conducted in two settings. About one-half of the students were trained in a local YMCA camp. The remaining students were trained in their local school setting. Evaluation of the trained students' grades in mathematics, English, and social studies over the school year showed that trained students performed significantly better in mathematics than a randomly selected control group of students. Evaluation of pre- and post-training standardized test scores in science and social studies showed that the trained students performed significantly higher on the science tests than did control students.

AS AN OUTGROWTH of the development of the measure of need for Achievement (10) and subsequent research dealing with the achievement motive, there has been developed an achievement motivation training (AMT) course. The training course is an effort to change the motivation tension system of the course participants. In essence, AMT is an answer to the question, "How does one go about motivating people?"

McClelland has called motives, "affectively toned associative networks," in the cognitive makeup of a person (12:322). The saliency of a motive in the cognitive network of a person can be measured by counting the number of associations belonging to the motive cluster. The associative network concerning competition with a standard of excellence has become the definition of the achievement motive, and the thought sampling technique of measuring achievement motivation is designed to tap fantasy associations clustered around competitive actions (10).

AMT is an attempt to increase the precision and, thus, the saliency of need for Achievement (*n* Achievement) thinking. The training, furthermore, is designed to give course participants an opportunity, in a series of game experiences, to examine the workings of their competitive strivings. McClelland (11), in a summary of *n* Achievement research, indicated that persons high in *n* Achievement act in certain ways: in performing challenging tasks where skill is involved, they take moderate risks; their strategy seems to be to get and use as much concrete feedback as possible; when approaching a challenge, the person high in *n* Achievement seeks out ways to take innovative action; and finally, those high in *n* Achievement thinking appear to search out positions of responsibility, namely, action areas where they can feel that their individual efforts in either group or personal goal-directed activity will make a difference. The term which best captures the essence of the achievement motive construct validity picture according to McClelland (11) is "entrepreneurship."

It is beyond the scope of this paper to describe in detail the AMT course. Various versions of AMT have appeared elsewhere (3;13;16;17). Generally, the course involves a minimum of four to five days of activities carried out in a live-in residential setting. The residential setting helps to create an atmosphere of total involvement away from the humdrum of everyday activity. The format for most of the game activities is performance in the activity followed by a discussion period in which each participant is encouraged to review his thoughts, strategies, and reactions in the activity.

Perhaps two examples of AMT inputs will help the reader visualize its nature. Course participants are first given the McClelland *n* Achievement test. They are then taught to score the test protocols (cf. 10). The intent is not to make expert scorers, but to make precise the definition of the achievement motive. Participants are encouraged to make jargon use of scoring categories in the discussion of course

activities. Theoretically, the principle behind teaching the scoring system is similar to that of psychotherapy. The more one verbalizes about an experience, the more the experience may become meaningful.

An example of the game activity is the Ring Toss Game. Each participant is given an opportunity to toss a number of rings over a peg while standing at any distance from the peg that he chooses. After several turns the results of each participant's efforts are recorded for all to see, and each is asked to discuss his or her strategy. The discussion readily reveals the moderate risk-taking strategy and strategies that avoid the challenge of the game. Those who stand close to the peg, a low-risk distance, are maximizing success possibilities but incur the good natured wrath of other participants as expressed in the comment, "Anyone could do it from there." Those who stand too far away from the peg, a caution to the wind or high-risk distance, seem to be hiding behind the comment, "I tried but no one could be expected to succeed from that distance." If successful, however, they reap the paradoxical benefit of either being considered very skillful or, more likely, being chided by the comment, "You were just lucky." The point is that moderate distances are those which offer the most challenge, and it is at moderate distances where those high in *n* Achievement have been found to stand (11). The Ring Toss Game and other similar experiences offer participants concrete opportunities to witness the workings of the achievement motive both in themselves and in others. The question, "Am I a person high in *n* Achievement or do I want to be?" is brought squarely into focus.

AMT, although principally used in the training of businessmen (13), has found its way into the educational system. Burris (1) launched the AMT effort by his initial attempt to improve the school performance of college students by counseling them using the *n* Achievement scoring system as a guideline. His results indicate that his counselees' grades improved. Kolb (9) found AMT effective in increasing the grades of bright tenth grade students from upper socioeconomic homes. Recently deCharms (4) has reported evidence of increased school performance of black sixth graders in St. Louis as a result of being taught by teachers who had experienced AMT and who structured some of their regular teaching units around the AMT experiences.

The present study was designed to test whether a group of teachers given a revised form of AMT could, in turn, by training students outside of their regular school activity, generate a measurable improvement in their academic performance. Simplified, the chain of reasoning was: AMT will lead to increased *n* Achievement thinking, which will lead to increased achievement behavior, which will result in improved academic performance. The study was mainly designed to test the hypothesis: AMT will have the effect of increasing the achievement motivation of students and lead to improvement in their academic performance.

Method

First, the study design involved training for a group of 24 teachers during the last two weeks of their summer vacations. Two training sessions of four and one-half days each were conducted. Each session was attended by twelve teachers, an expert trainer and the author.[2] The teachers then trained students. Students were given AMT in two settings. One group of students recruited from the schools participating in the study was given the training in a live-in camp setting, a local YMCA camp in the area. Another group of students was given AMT in their local school setting. A third group was given no training. Students were selected from both the eighth and tenth grades. The training of students was conducted on four weekends during their fall school semester.

Subjects

Officials from three high schools and five intermediate schools agreed to cooperate in the study. The schools were located in northern, central, and southern San Mateo County, California.

The first step in the selection of students was to identify average ability, low-achieving students from each cooperating school. The cumulative records of the schools included intelligence scores for all students. For each school a regression coefficient was computed based on an average of the students' grades in English, social studies, and mathematics and their intelligence test scores. From each school, those students with intelligence scores between 85 and 120 who were going into the eighth and tenth grades, and who were not achieving up to their predicted levels as indicated by the overall achievement of the beginning eighth and tenth graders in their school were identified as low achievers.[3] From the low-achieving group of each participating school a control group was randomly selected. Following the control group selection, volunteers for AMT were recruited from each of the participating schools. Sixty-four eighth grade students and 78 tenth grade students volunteered (average age: eighth grade—13 years 10 months, tenth grade—15 years 9 months; average IQ: eighth grade—105, tenth grade—104). The average IQ for the tenth grade control students was 105 (n = 45). The average IQ for the eighth grade control students was also 105 (n = 30). The difference between experimental and control students' IQ was not statistically significant.

Students who volunteered for AMT might be considered more motivated to start with. However, nearly everyone who was asked to participate accepted. Those who refused gave other commitments (work, athletics, etc.) as reasons for their refusal. Initially, at least, differences in performance after training could not be due to differences in desire to participate in AMT.

Procedure

There were six separate experimental groups for both eighth and tenth grade. Three groups from each grade were

trained in a camp setting and three groups from each grade were trained in their local schools. Two trained teachers were assigned to each of the experimental groups. Students who volunteered for the AMT course were randomly assigned to the camp or school setting. Students assigned to the camp setting were bussed to the training site the Saturday morning of the training weekend and returned Sunday evening. Students assigned to the school setting spent Saturday and Sunday of the training weekend at their respective school, but returned home the Saturday night of the training weekend. The only difference between the camp and campus training format was that the camp group did experience some of the live-in impact of training. Obviously, the experimental arrangement was designed to see if the live-in impact was an essential part of the training format.

Prior to AMT, the experimental and control students were given the Stanford Achievement Test, Form W, science and social studies tests, Part A. In the late spring of the following semester, Form X of the same test battery was administered. Experimental and control students were brought together in a single testing session and told the tests were part of the school testing program, but that the scores would not be shown on their school records. The scores used in the analysis of data were residualized gain scores computed from the regression of Form X on Form W. Using the initial Form W scores and the regression coefficient, a predicted score was determined for each student. The predicted score was subtracted from the student's actual Form X score, and then 50 points were added to eliminate negative scores.

Grade point average (GPA) scores were manipulated in the same manner using June grades prior to training and June grades after training in English, mathematics, and social studies. The GPA scores were residualized in the same manner as the standardized test scores.

Table 1.—Effect of AMT on Gain Scores for Students Who Attended at least Three of the Four Training Sessions

	Social Studies Test			Science Test	
Training Group	N	$\bar{X}$	SD	$\bar{X}$	SD
Campus	41	49.6	5.8	50.8	6.2
Camp	43	50.2	5.1	51.3	6.1
Control	49	49.3	5.0	46.9	6.9

Table 2.—Analysis of Variance of Science Gain Scores for Students Who Attended at least Three of the Four Training Sessions

SOURCE		df	MEAN SQUARE	F
GRADE LEVEL	(A)	1	56	1.35
SCHOOL (NESTED)	(B)	4	153	3.70*
TRAINING	(C)	2	208	5.03*
AxC		2	12	<1
AxBxC		8	45	1.08
ERROR		117	41.3	

*$p < .01$

Attendance figures indicate that 34 of the original 136 students who volunteered and who attended the beginning training sessions dropped out of training before the third or fourth session. That drop-out rate (25%) is not astonishing when one takes into account that students were asked to volunteer their weekends for training. It is obvious, however, that cooperative attitude reflected in willingness to attend sessions was a source of error.

The significant school nested in grade level main effect shown in Tables 2 and 5 will not be discussed. This was as expected, due to the different composition of student body of the schools selected for the study, and is of little theoretical interest.

Results

A 3 x 3 x 2 analysis of variance was applied to evaluate the main effects and interactions among the factors training, grade level, and school nested in grade level. Table 1 shows a comparison of the gain score means for experimental students who attended at least three sessions with the gain score means for the control students. Table 2 shows the analysis of variance for the science scores presented in Table 1. Inspection of the mean scores for social studies test scores indicates no support for the predicted training effect. The science scores were in the predicted direction. Inspection of Table 1 shows that the trained groups scored higher on the science tests than did the control group, and Table 2 shows that the differences in performance exceeded chance expectancies.

Table 3 shows the June grades of those students who attended three training sessions. Inspection of the means shows that only the mathematics grades were in the predicted direction. Table 4 shows the residualized gain score for the mathematics grades. Table 5 shows the analysis of variance for the gain scores in mathematics. Inspection of Table 5 shows that the trained groups' gain scores exceeded chance expectancies.

Discussion

Certainly, the results indicate only marginal support for the hypothesized effect. To some extent, however, that was expected. The AMT course was in many of its facets less than ideal. Teacher trainers were considerably less than expert in their training skills. As it was their first training effort, the "bugs" were not yet out of the system. Because the training itself was done on weekends, the total involvement so important for the self-study and interpersonal support facets of the training was weakened.

Table 3.—June Grades for Students Who Attended at least Three of the Four Training Sessions

Training Group	English N	English $\bar{X}$	Social Studies $\bar{X}$	Mathematics $\bar{X}$
Campus	45	5.5	5.9	5.5
Camp	47	5.8	5.6	6.1
Control	69	5.8	6.2	4.9

Table 4.—Effect of AMT on Mathematics Grades Gain Scores for Students Who Attended at least Three of the Four Training Sessions

Training Group	N	Mathematics Grades $\bar{X}$	Mathematics Grades SD
Campus	45	52.3	20.3
Camp	47	59.6	22.6
Control	69	47.0	19.8

Table 5.—Analysis of Variance of Mathematics Grade Gain Scores for Students Who Attended at least Three of the Four Training Sessions

SOURCE		df	MEAN SQUARE	F
GRADE LEVEL	(A)	1	154	<1
SCHOOL (NESTED)	(B)	4	1207	2.76*
TRAINING	(C)	2	1690	3.87*
AxC		2	395	<1
AxBxC		8	679	1.78
ERROR		143	436	

$p < .05$

The two-step process of training, that is, teacher being trained and then training students, in spite of the weaknesses in the arrangement, was deemed necessary to show the value of AMT for school use. Previous in-school use of AMT involved the direct contact between the expert trainer and students. The present study and the one conducted by deCharms (4) were conducted to show that training skills could be put in the hands of teachers after minimal exposure to expert training. If they, then, could use those skills to influence student academic performance, the initial foundation for curriculum innovation would be laid.

The results in the present study, the data configuration when viewed in the context of the *n* Achievement construct validity picture, does make sense. Kagan and Moss (8) found a positive correlation between *n* Achievement in middle childhood and skill at constructional activities. McClelland and Winter (13) pointed to the fact that constructional activities furnish the concrete feedback concerning performance that those high in achievement motivation desire. For example, the difference in feedback in building a radio set and handing in a theme is related to the need-for-feedback predisposition of the person high in achievement motivation. In the radio set case, the completion of the last solder signals the turning-on of the set. It either plays or does not. In the theme case, of course, the feedback is not immediately forthcoming, and, perhaps of more importance, is not as completely controlled by the person handing in the theme as it is in the radio set construction. Both science scores and mathematics scores, it could be argued, are sensitive to the person's need for concrete feedback. The science test items are a sample of constructional activity-related knowledges. Mathematics as an exercise permits immediate feedback about performance for anyone conscientious enough to check his answers. Apparently the training, where effective, was particularly appealing to those students who preferred to operate in task areas where they could gain immediate concrete feedback about their performance.

Most researchers involved with AMT have been hard pressed to describe exactly what changes among trainees take place in training. Personal experiences with AMT have suggested that the training cultivates a person's feeling of personal effectiveness. McClelland and Winter (13) in discussing the reactions of businessmen to training have described the training effect as an upsurge of certainty that one can control one's destiny, a feeling of personal efficacy. DeCharms (2), building on Heider's (6) notions of internal locus of control, has singled out the Origin-Pawn dimension as pertinent to the discussion of the achievement motive and human motivation in general. The Pawn feels pushed around by controls from outside; the Origin feels that he himself is in control, that is, he is controlled from within. Rotter (15) has described facets of the Origin-Pawn dimension in his discussion of internal and external control of reinforcement. These motivational variables (or single variable) seem to be a facet of the instinct to master (7) and effectance motivation (18). In other words, there is apparently a theoretical affinity between *n* Achievement and a basic human desire to feel personally in the driver's seat of one's vehicle of destiny. Apparently AMT creates a series of peak experiences which arouse in trainees a sense that they can do more to control their destiny. Couched in the study of achievement strategies, that feeling, at least for the students involved in the present study, apparently manifested itself in subject matter areas which were sensitive to the need-for-feedback and Origin predisposition engendered by AMT.

Several implications of AMT are of concern to those involved in this type of applied research. One is the magnitude of change that can be effected as a result of AMT. Upon examination of the changes in grades represented by the coded scores reported on Table 1, they are seen to be of the D+ and C– to C magnitude. One wonders about the expenditures of training energy in relation to expected change. In change efforts which deal with motivation,

however, it may be enough to show that when what has been a trend in the Ss past schooling—a trend toward falling increasingly behind his peers in achievement—is changed to one where the gap between his and peer achievement is narrowed, the effort is worthwhile.

A second concern deals with marginality of results. Most of the AMT projects have not shown remarkable differences in the behaviors of experimental and control subjects (14). One would certainly like to see more convincing support in each individual research effort. Marginality, however, may be the plight of those involved in applied research. Applied research places limits of administrative expediency on the control of variables. Furthermore, variables central to a person as a person, namely, attitudes, motives, and other belief systems in general, are simply not changed in drastic ways. If they could be changed drastically, in a one shot effort, they would not be central to a person's "core" functioning. What is encouraging in the construct validity picture surrounding AMT is that a series of projects, all of which have produced evidence supporting the effectiveness of AMT as a process to produce motivational change, have been carried out.

FOOTNOTES

1. The present study was supported by Research Grant OEG-3-7-703306-4256, Project Number 67-3306 of the Elementary and Secondary Education Act, Title III.

2. The expert trainer was Monohar Nadkarni, the trainer involved in training Indian businessmen for the studies reported by McClelland and Winter (13).

3. The phrase "low achiever" is purposely used to avoid the connotation of underachievement. The underachieving syndrome (5) is a complicated issue involving rebelliousness, resentment of criticism, anticipation of failure, and other personality predispositions to which the mathematical definition of low achievement used in the present study was not sensitive.

REFERENCES

1. Burris, R. W., "The Effect of Counseling on Achievement Motivation," unpublished doctoral dissertation, University of Indiana, Bloomington, 1958.
2. DeCharms, R., *Personal Causation,* Academic Press, New York, 1968.
3. DeCharms, R., "From Pawns to Origins: Toward Self-Motivation," in G. S. Lesser (ed.), *Psychology and Educational Practices,* Scott, Foresman, Chicago, 1971, pp. 380-407.
4. DeCharms, R., "Personal Causation Training in the Schools," *Journal of Applied Social Psychology,* 2:95-113, 1972.
5. Fine, B., *Underachievers: How They Can Be Helpted,* E. P. Dutton, New York, 1967.
6. Heider, F., *The Psychology of Interpersonal Relations,* Wiley, New York, 1958.
7. Hendrick, I., "Instinct and the Ego during Infancy," *Psychoanalytic Quarterly,* 2:33-58, 1942.
8. Kagan, J.; and Moss, H., *Birth to Maturity: A Study in Psychological Development,* Wiley, New York, 1962.
9. Kolb, D. A., "Achievement Motivation Training for Underachieving High School Boys," *Journal of Personality and Social Psychology,* 2:783-792, 1965.
10. McClelland, D. C.; Atkinson, J. W.; Clark, R. A.; and Lowell, E. L., *The Achievement Motive,* Appleton-Century-Crofts, New York, 1953.
11. McClelland, D. C., *The Achieving Society,* D. Van Nostrand, New York, 1961.
12. McClelland, D. C., "Toward a Theory of Motive Acquisition," *American Psychologist,* 20:321-333, 1965.
13. McClelland, D. C.; and Winter D. G., *Motivating Economic Development,* Free Press, New York, 1969.
14. McClelland, D. C., "What Is the Effect of Achievement Motivation Training in the Schools?" *Teacher College Record,* 74:130-148, December 1972.
15. Rotter, J. B., "Generalized Expectancies for Internal versus External Control of Reinforcement," *Psychological Monographs,* vol. 80, no. 609.
16. Ryals, K. R., "An Experimental Study of Achievement Motivation Training as a Function of the Moral Maturity of Trainees," unpublished doctoral dissertation, Washington University, St. Louis, 1969.
17. Shea, D.; and Jackson K., "Motivation Training with Teachers—A Description," unpublished paper, Washington University, St. Louis, 1970.
18. White, R. W., "Ego and Reality in Psychoanalytic Theory," *Psychological Issues,* vol. 3, no. 11.

TEACHER CLASSROOM MANAGEMENT SKILLS AND PUPIL BEHAVIOR

WALTER R. BORG
Utah State University

PHILIP LANGER
University of Colorado

JEANETTE WILSON
Northern Colorado Bureau of Cooperative Services

ABSTRACT

An experimental group of 20 inservice elementary teachers was trained using the Utah State University Classroom Management Protocol Modules, and compared before and after training with a control group of 9 teachers. Although the experimental teachers received more favorable post-training scores on all 13 classroom management behaviors covered in the modules, the differences were generally small and nonsignificant. The level of work involvement and deviant behavior of pupils of the experimental group teachers was also compared before and after the teachers had been trained. In recitation situations, pupil work involvement increased and deviant behavior decreased significantly. In seat work situations, pupil work involvement increased significantly, but no significant changes occurred in deviant behavior.

THE PURPOSE OF this study was to determine whether the Utah State University Protocol Modules that are designed to improve the classroom management skills of elementary teachers brought about significant changes in the teacher use of these skills and also changed the amount of on-task and disruptive behavior of pupils in their classes.[1]

Specific behaviors taught by the U. S. U. Classroom Management Modules were drawn primarily from a correlational study carried out by Kounin (5). In this study, Kounin collected videotapes in 49 elementary school classrooms. He identified two pupil performance criteria, work involvement and deviant behavior. Eight pupils were selected for observation in each classroom, four boys and four girls. Each child was scored for work involvement and deviant behavior every 12 seconds. Work involvement was scored in three categories: (1) definitely in the assigned work; (2) probably in the assigned work; (3) definitely out of the assigned work. Deviance was also coded by Kounin into three categories: (1) not misbehaving; (2) engaging in mild misbehavior; (3) engaging in serious misbehavior. Teacher behaviors related to classroom management were scored by a different group of observers in Kounin's study. The observational procedures differed for the different teaching behaviors with some tallied on 6-second intervals, while others were rated at 30-second intervals. The observational procedures used in the study reported herein differed from those used by Kounin and will be described later in this paper. Table 1 summarizes correlations obtained by Kounin between teacher behavior and pupil work involvement and deviant behavior for those variables that were covered in the U. S. U. Protocol Modules.

The following hypotheses are proposed:

1. There will be no significant difference between the adjusted post-treatment performance of trained and untrained teachers on the teacher behaviors covered in the U. S. U. Classroom Management Modules.
2. For teachers trained in the modules there will be no significant difference in the frequency of pupil work involvement and deviant behavior in their classes before and after training.

Procedures

Subjects

Ss in this research consisted of 29 inservice elementary school teachers employed in the Denver area. Twenty of these teachers were trained in a course in which the U. S. U. Classroom Management Protocols were employed. The other nine teachers constituted a comparison group that did not receive the classroom management training. These teachers were drawn for the most part from the same schools as the experimental group teachers. However, teachers were not assigned to the two treatments randomly.

Treatments

Teachers in the experimental group were enrolled in an extension course in classroom management offered by the University of Colorado and taught by Dr. Jeanette Wilson.

Table 1.–Correlations between Teacher and Pupil Behavior as Reported by Kounin

Teacher Behavior	Pupil Behavior: Recitation – Work Involvement	Recitation – Freedom from Deviancy	Seatwork – Work Involvement	Seatwork – Freedom from Deviancy
1. Withitness	.615	.531	.307	.509
2. Transitions	.601	.489	.382	.421
3. Group Alerting	.603	.442	.234	.290
4. Learner Accountability	.494	.385	.002	-.035

Table 2.–Teacher Behaviors Emphasided in the U. S. U. Classroom Management Protocol Modules

TRANSITIONS

1. *Stimulus Boundedness:* The teacher is deflected from the main activity and reacts to some external stimulus that is unrelated to the on-going activity, vs. *Delayed Response:* The teacher delays responding to an unrelated stimulus until a natural break occurs in the classroom activity.
2. *Thrust:* The teacher bursts in suddenly on the children's activities in such a manner as to indicate that her own intent of thought was the only determinant of her timing and point of entry, vs. *Timely Interjection:* The teacher introduces information in a manner which minimizes interruption to the students' activity.
3. *Flip-Flop:* The teacher starts a new activity without bringing the original activity to a close and then returns to the original activity, vs. *Smooth Transition:* The teacher fully completes one activity before moving on to the next.

LEARNER ACCOUNTABILITY

1. *Goal Directed Prompts:* The teacher asks questions which focus on the student's goal by asking about his *work plans* or *work progress.*
2. *Work Showing:* The teacher holds students accountable for their work by having them *show work* or *demonstrate* skills or knowledge.
3. *Peer Involvement:* The teacher involves students in the work of their peers by having them respond to another student's recitation or work activity.

GROUP ALERTING

1. *Questioning Technique:* The teacher frames a question and pauses *before* calling on a reciter (QT+), rather than naming the reciter and *then* giving the question (QT–).
2. *Recitation Strategy:* The teacher calls on reciters at random (RS+) rather than calling on them in a predetermined sequence (RS–).
3. *Alerting Cues:* The teacher alerts nonperformers that they may be called on (AC).

WITHITNESS

1. *Desist:* The teacher demonstrates Withitness by telling students to stop the deviant or off-task behavior. In order to be effective, the desist must be directed at the student who initiated the deviant behavior and must be administered before the deviant behavior spreads or becomes more serious. It must be timely and on target (D+). If the desist is not timely or on target, it is a negative desist referred to as (D–).
2. *Suggest Alternative Behavior:* When deviant behavior occurs, the teacher diverts the disruptive or off-task student by suggesting that he engage in an alternative behavior.
3. *Concurrent Praise:* The teacher avoids direct confrontation with a student who is displaying deviant or off-task behavior by concurrently praising the non-deviant or on-task behavior of other students.
4. *Description of Desirable Behavior:* The teacher describes or has the off-task student describe the desirable behavior which the student usually exhibits or should exhibit in place of the on-going deviant or off-task behavior.

This course extended over a period of ten weeks and met for three hours per week. The course content consisted primarily of the four Classroom Management Protocol Modules developed at Utah State University. Each module was concerned with a major concept adapted from Kounin's research. In completing each module the teacher focused on three or four specific behaviors that could be used in the classroom to apply the general concept. These concepts and the specific related behaviors are listed and defined in Table 2.

Each of the U. S. U. Protocol Modules consists of a Student Guide, a Protocol Film, and a set of evaluation materials. In completing a module, the teachers being trained went through the following steps:

1. Scan the *Learning Sequence.* This gives the learner a step-by-step outline of what he will do.
2. Read the module objectives, a description of the concept and the three specific teacher behaviors to be used to apply the concept in a classroom. In some modules, desirable and undesirable behaviors are contrasted.

3. Complete the *Recognition Practice Lessons.* These are transcripts made from classroom audiotapes. The learner must identify instances when the teacher used the behaviors being learned and determine which behavior was used.

4. View the *Protocol Film* and identify instances when the teacher in the film used the behaviors covered in the module. This film also provides a model for the learner.

5. Take a performance test, the *Recognition Test,* designed to measure the learner's ability to recognize classroom applications of the teacher behaviors and discriminate between applications and non-applications.

6. Plan a brief lesson designed to practice the Classroom Management behaviors. The teacher teaches this lesson in his or her own class and records it on audiotape.

7. Replay the lesson with another teacher who is participating in the training, record on a tally sheet use of the behaviors practiced, and discuss.

Teachers in the experimental group were aware of the fact that they were involved in a project aimed at evaluating the U. S. U. Classroom Management Modules. The investigators emphasized that the study was aimed at evaluating the modules and determining how they could be improved, and not aimed at evaluating the *Ss* as teachers.

The control group teachers received no training in classroom management during the period of the study. Observations of their classroom management behavior were made during the same time that pre- and post-observations were being made of the experimental group teachers. The control group teachers were aware that they were participating as control group members in a study aimed at evaluating a new course of study. Several days prior to the observation all teachers were given identical instructions, which included a list of the specific teacher behaviors that would be observed in their classroom.

Teacher Observations

Two observers were trained to collect data on teacher performance. Training consisted of studying the protocol modules, practicing, and meeting with Dr. Langer to clarify the procedure and resolve problems. The observation of teacher performance involved tallying the frequency with which teachers exhibited 19 specific behaviors related to classroom management. Of these 19 behaviors, it was found that the three positive *Transition* behaviors were extremely difficult to observe. It is quite easy for the observer to detect a teacher "thrust" that occurs in the classroom situation. However, a situation in which a teacher could use a "thrust" and does not is very difficult to detect. Therefore, data obtained on the three positive *Transition* behaviors was not employed in the analysis. The negative *Transition* behaviors were found to occur at a low frequency in the classrooms in this study. Therefore, these three behaviors were combined to yield a total negative *Transition* score. The *Withitness* behaviors, except for "positive desists" (D+), also occurred with very low frequencies during the pre-training observations. Tallies of positive and negative "recitation strategy" also occurred at a very low level probably because of an unsatisfactory operational definition. Since the observational frequencies for most teachers on "recitation strategy" was zero, no reliability coefficient was computed.

In order to compute the seven reliability coefficients shown in Table 3, the two observers independently observed the same teacher during the same time span on ten occasions. These observations were all conducted during the pre-training observational period. The length of these observations ranged from 40 to 50 minutes, with a mean observation time of 47 minutes. Rank difference correlations were computed for each teacher behavior in order to obtain inter-observer reliability coefficients. It will be seen in Table 3 that these coefficients range from .60 for goal-directed prompts to .97 for positive questioning technique (QT+). Since the specific behaviors under *Transitions* and *Withitness* occurred at too low a frequency to produce reliable scores, it was decided not to work with subscores in the analysis but instead to combine the subscores under each of the four major categories and carry out analysis on these four major category scores.

Although the inter-observer reliability coefficients obtained on the pre-training teacher observations were satis-

Table 3.—Inter-observer Reliability on the Classroom Management Variables*

Variable	Rho
1. Transitions (total negative-i.e. sum of SB-, T-, FF-)**	.84
2. Learner Accountability	
a. Goal Directed Prompts (GDP)	.60
b. Work Showing (WS)	.96
c. Peer Involvement (PI)	.90
3. Group Alerting	
a. Positive Questioning Technique (QT+)	.97
b. Positive Recitation Strategy (RS+)**	
c. Alerting Cues (AC)	.93
4. Withitness (total positive-i.e. sum of D+, SAB, CP AND DDB)**	.88

* Rho correlations based upon ten observations made before teachers were trained. Mean observation time, 47 minutes.
** Frequencies of subscores very low with zero entries for several teachers; no correlations computed on subscores.

factory, because of the low frequencies of many of the behaviors, it was decided to attempt to collect two hours of observational data on each teacher after training rather than the 50 minutes of observation conducted during the pre-training observational period. The post-training teacher observations were carried out by the same two observers, and ranged in length from 105-120 minutes. Since the pre-observational reliabilities had been satisfactory, joint observations in the same classrooms were not carried out for any of the post-treatment observations, and inter-observer reliabilities were not computed. In order to make the pre- and post-frequencies comparable, all observational frequencies were converted to a 120-minute base.

Pupil Observations

Pre-post pupil observations were conducted only in the rooms of experimental group teachers. The task of the pupil observers was somewhat easier than the teacher observers'. The pupil observer stationed herself at one side of the room near the front so as to be able to see the faces of all children in the room. The observer started at one corner of the room and observed a child for a period of about eight seconds. The observer then tallied the child into one of five pupil behavior categories; that is, (1) definitely involved in class work; (2) probably involved; (3) definitely off-task; (4) mildly deviant; (5) seriously deviant. The observer then looked at the next child for a period of eight seconds and continued this process until she had viewed and tallied the behavior of each child in the classroom. This observational technique resulted in the observer's tallying the behavior of each child once every four to five minutes. The observer was also instructed to be alert for any instances of seriously deviant behavior and to record these even if watching a different child at the time. Since seriously deviant behavior usually attracts a good deal of attention, it was believed that the observers would probably pick up nearly all cases of seriously deviant behavior. On the other hand, the observer was instructed to record the other four pupil behaviors only for the specific child being watched during the particular eight-second interval.

It was decided that it would be desirable to have approximately 50 observations of each child. Since the observer would obtain approximately 12 observations on each child per hour, a total observational period of four hours was selected for the pupil observations. Pupil observations were usually collected on four or five different occasions. The length of the individual observational periods varied considerably in order to fit into various situations that occurred in different classrooms during the school day. For example, if an observer entered a classroom, observed 30 minutes and then the class left the room for a meeting in the auditorium, the observer would terminate the observation at 30 minutes. In spite of this variation in the length of individual observation sessions, however, the total observation time for each teacher was reasonably close to four hours.

Four observers were trained to collect the pupil observational data. During the pre-training observations, pairs of these observers were scheduled to observe a total of 12 observational sessions in the same classrooms. When an attempt was made to obtain inter-observer reliabilities from these 12 sessions, however, a number of problems were encountered. The most serious problem was the inability to devise a system that would keep the observers in phase, i.e., observing the same child during the same eight-second time span. Since a child may be on task one minute and off task the next, comparisons of pupil behavior for the two observers would tend to underestimate the inter-observer reliability. As a result of this and other problems, satisfactory inter-observer reliability estimates on the five pupil behavior variables were unable to be obtained.

The present authors are currently carrying out a study in which they have managed to solve this problem. In the current study, all four pupil observers have observed in the same classrooms during the same 30-minute periods on 11 different occasions. In order to keep the observers in phase, all observers started at a given point in the classroom and observed the pupils around the classroom in a predetermined order. The head observer held up her pencil while observing the first child, then put the pencil down to tally the child's behavior, and raised the pencil while observing the second child. Other observers followed these pencil signals in order to stay in phase with the head observer. Inter-observer reliability coefficients in this study ranged from .93 to .99. These data should be considered as only a rough indication of probable inter-observer reliability in the current study.

Pupil performance observations were carried out only in the classrooms of the experimental group teachers. Data were collected separately for times when the class was involved in recitation and times when the class was involved in seat work. During periods when part of the class was involved in recitation and part in seat work, the time was divided equally between the two categories.

Although the average observation time for both pre- and post-observations was four hours each, the proportion of recitation time vs. seat work time differed from teacher to teacher. On the pre-training observation, the mean pupil observation time for recitation was 140 minutes. The mean observation time for seat work on the pre-training observation was 96 minutes. For the post-treatment observations, these means were 141 minutes for recitation and 102 minutes for seat work.

In order to make the data comparable from classroom to classroom, the actual frequencies of pupil behavior that were recorded during recitation were multiplied by a factor obtained by dividing 140 minutes by the actual observation time. This result gives an estimate of the frequency of each pupil behavior tally that would have been obtained if the recitation time for all classrooms had been 140 minutes. Since this was the mean recitation time,

Table 4.—Adjusted Final Means on Classroom Management Variables

Variable	Exp. Adj. final mean	Con. Adj. final mean	Adj. F
1. Stimulus Boundedness (SB-)	.11	.27	NS
2. Thrust (T-)	.04	.24	NS
3. Flip-Flop (FF-)	.05	.27	NS
Neg. Transitions Total	.20	.78	NS
4. Goal Directed Prompts (GDP)	4.63	2.75	NS
5. Work Showing (WS)	14.27	8.18	NS
6. Peer Involvement (PI)	3.37	1.77	NS
Learner Accountability Total	22.27	12.70	.05
7. Positive Questioning Techinque (QT+)	24.93	20.66	NS
8. Positive Recitation Strategy (RS+)	1.98	1.35	NS
9. Alerting Cues (AC)	2.13	.49	NS
Group Alerting Total	29.04	22.50	NS
10. Positive Desists (D+)	4.73	4.25	NS
11. Suggest Alternative Behavior (SAB)	2.86	2.85	NS
12. Concurrent Praise (CP)	2.17	1.78	NS
13. Describing Desirable Behavior (DDB)	4.23	3.15	NS
Withitness Total	13.99	12.03	NS
Composite of 10 favorable behaviors	65.30	47.23	NS

Table 5.—Changes in On-task and Deviant Behavior in Classrooms of Teachers Who Completed the U. S. U. Classroom Management Modules

Pupil Behavior	Pre-training Mean	%	Post-training Mean	%	t*
Recitation (140 minutes)					
1. Definitely involved in classwork	541.8	80.3	686.2	88.6	1.89
2. Probably involved in classwork	127.1		52.3		3.47
3. Definitely off task	114.0	13.6	71.3	8.5	2.24
4. Mildly deviant behavior	48.4	5.8	22.9	2.7	1.94
5. Seriously deviant behavior	1.6	---	.2	---	2.20
Seat Work (100 minutes)					
1. Definitely involved in classwork	393.1	80.1	514.8	82.9	2.48
2. Probably involved in classwork	51.8		40.0		1.11
3. Definitely off task	75.7	13.6	71.9	10.7	.26
4. Mildly deviant behavior	33.4	6.0	40.6	6.0	.59
5. Seriously deviant behavior	.9	---	1.4	---	.04

*t > 1.74 is significant at .05 level; t > 2.57 significant at .01 level using one-tailed test.

multiplying the actual scores by this factor did not change the overall mean for each pupil behavior.

For seat work, the pupil behavior scores in each classroom were multiplied by a factor obtained by dividing 100 minutes by the actual number of minutes of seat work observation. Again, since the mean seat work observation time was very close to 100 minutes, the result of multiplying this factor was to provide an estimate of each pupil behavior frequency that would have occurred if seat work observation in all classrooms had in fact been 100 minutes long.

Results

In order to determine whether teachers trained in the U. S. U. Classroom Management Modules made significantly greater gains on the 13 specific classroom management behaviors than control group teachers, an analysis of covariance was carried out on each behavior, on each of the four behavioral categories, and on a composite score made up of the ten desirable behaviors. For each analysis the teachers' pre-treatment performance was used as the covariate and post-treatment performance was used as the dependent variable. The adjusted final means for the experimental and control teachers are given in Table 4. It will be noted that all of the adjusted final means favor the experimental group. However, for the most part, differences between the experimental and control groups were very small. There were no significant differences between the adjusted final means on any of the 13 specific behaviors covered in the four modules. Of the four composite module scores, the adjusted final mean score of the trained teachers was significantly higher than the untrained teachers only in *Learner Accountability*. The composite score made up of ten favorable teacher behaviors, omitting the three negative transition behaviors, shows a somewhat higher frequency for the trained teachers. However, the difference in adjusted final means was not statistically significant. In looking over the adjusted final means for both groups, it is clear that most of the 13 behaviors occurred at low frequencies during the 120-minute observational period. Only "work showing," "positive questioning technique," and "withitness" occurred with moderate frequency. These frequencies suggest that two hours may not have provided a valid indication of teacher use of the behaviors taught. It will be recalled that Kounin's data on teacher behavior were collected over an entire school day. A replication of the study reported herein is currently underway, and in this new study pre- and post-observational data are being collected for each teacher over an entire school day.

The results of the current study, however, generally failed to show significant differences between the adjusted post-mean performance of teachers in the experimental and control groups. Therefore, the null hypothesis cannot be rejected, and we must conclude that the U. S. U. Classroom Management Modules did not bring about significant changes in the specific behaviors taught.

Pupil Behavior

Data on pupil behavior were collected before and after training only in the classes of the experimental group teachers. Pre-training observations were made in late January and early February, while post-training observations were made in late April and early May. The frequency of each of the five categories of pupil behavior before and after the teachers were trained were checked for significance using the *t*-test. The results of the pre and post pupil observations under recitation and seat work conditions are given in Table 5.

In the recitation condition, it will be noted that the number of pupils definitely involved in seat work was significantly higher on the post-training observations, while the number of students probably involved in class work was significantly lower. The lesser use by the observers of the "probably involved" category during the post-training observations could have reflected the greater amount of observational experience they had had at that time. When these two categories are lumped together, it will be noted that 80.3% of the pupils were definitely or probably involved in class work during the pre-training observation as opposed to 88.6% during the post-training observation. The number of pupils definitely off-task dropped significantly between the pre- and post-training observations. Only 8.5% of the students were observed as being definitely off-task during the post-training observation, as compared to 13.6% during the pre-training observation. The frequency of mildly deviant behavior on the post-training observation dropped to less than half of the pre-training frequency.

It should be remembered that all frequencies given in Table 5 represent about 1/30th of the frequencies that would have occurred if all pupils had been observed continuously at a rate of 6 times per minute over the 140-minute observation. If our observational data obtained by viewing pupils in sequence is representative of the data that would be obtained if they had all been viewed continuously, then it can be estimated that the mean number of mildly deviant acts that would occur in the typical classroom of 30 pupils in an hour would have been 662 before the teachers were trained and 294 after the teachers were trained. Put another way, such acts would occur about ten times per minute during recitation in the pre-training classrooms and five times per minute in the post-training classrooms. Since most teachers believe that deviant behavior occurs much more frequently near the end of the school year, our results on mildly deviant behavior might have been even more striking if the post-observations had not occurred in late April and early May. This possibility could, of course, have been checked if pupil performance data had been obtained on the control group classrooms.

The frequency of seriously deviant behavior during recitation also dropped significantly between the pre- and post-observations. However, this behavior occurred at a

very low frequency rate in the classrooms studied. The reader will recall that observers were to record mildly deviant behavior only when it occurred during the time they were watching a given child. In other words, if the observer were watching Child A during a given eight-second interval and noticed mildly deviant behavior on the part of Child B during this time, the Child B behavior would not be recorded. In contrast, since seriously deviant behavior is usually intrusive enough to attract the attention in the classroom and since we expected from Kounin's research that it would be a low-frequency behavior, the observers were instructed to record acts of seriously deviant behavior even if they were committed by a child other than the one the observer was watching during a given eight-second interval. Therefore, the frequencies of seriously deviant behavior would probably not go up substantially even if all children in the classroom were continuously observed.

Seat Work

Pupil behavior data were recorded separately for recitation and seat work situations. During the four hours of pupil observation, approximately one hundred minutes of this observation took place during seat work activities in the average classroom. It will be noted in Table 5 that the number of pupils observed who were definitely involved in seat work increased significantly between the pre- and post-training observations. The percentage of observation in the "definitely" plus "probably involved" categories increased from 80.1% during the pre-observation to 82.9% during the post-observation. Frequencies of the other four pupil behaviors did not change significantly between the pre- and post-observations, although the percentage of definitely off-task behavior was reduced from 13.6% to 10.7%. Kounin's research found that the relationship between the teacher classroom management behaviors and pupil work involvement and deviant behavior was generally smaller for the seat work condition than during recitation. Nine of the thirteen specific behaviors covered in the U. S. U. Classroom Management Protocol Modules are aimed at establishing a classroom environment in which off-task and disruptive behavior is less likely to occur. The other four behaviors (the *Withitness* behaviors) are designed to provide the teachers with a means of responding to deviant pupil behavior when it does occur. Since the nine preventative behaviors can be used much more effectively by the teacher during recitation, it is not surprising that most of the significant changes in pupil behavior occurred during the recitation situation.

Discussion

It could be hypothesized that the significant improvements in pupil behavior obtained during recitation were due to a change in the rater's interpretation of the pupil behavior categories. Such subtle changes in observer frame of reference can occur in observational studies particularly when two sets of observations are separated by a period of time. The pupil observers did receive refresher training, however, in order to reduce the likelihood of this happening. Also, if the changes obtained during recitation were in fact due to changes in observer frame of reference, then changes of about the same magnitude should have occurred during the observation of pupils doing seat work. Since such changes did not occur in four of the five pupil behavior categories, it seems safe to conclude that the recitation data probably represent real changes in pupil behavior.

In this study the link that was found between the U.S.U. Classroom Mangement Modules and changes in pupil behavior would have been much stronger if significant changes in teacher performance had also been obtained. The failure to obtain significant changes in teacher performance could have been due to any combination of several factors. One possibility, of course, is that the modules simply failed to change the behavior of the teachers. However, since the authors have used a similar instructional model to change specific teacher behaviors in previous studies and since these studies have usually produced large changes between pre- and post-teacher behavior, this interpretation might not be correct (1, 2, 3; 4).

Another possibility is that since most of the classroom management behaviors covered in these modules occur at rather low frequencies, this study's observational time of approximately one hour pre and two hours post was inadequate to provide a representative sample of the teacher's use of the specific behaviors. It is also possible that a fatigue factor operated during the post-observations, which were twice as long as the pre-observations, which could have resulted in the observers' being less alert to teacher behaviors during the second hour of observation.

FOOTNOTE

1. The U. S. U. Classroom Management Protocol Modules may be purchased or rented from the National Resource and Dissemination Center, Division of Educational Resources, University of South Florida, Tampa 33620.

REFERENCES

1. Borg, Walter R.; Kelley, Marjorie L.; Langer, Philip; and Gall, Meredith, *The Minicourse–A Microteaching Approach to Teacher Education,* Macmillan Educational Services, Beverly Hills, Calif., 1970, 256 pp.
2. Borg, Walter R., "The Minicourse as a Vehicle for Changing Teacher Behavior: A Three-year Followup," *Journal of Educational Psychology,* 63:572-579, December 1972.
3. Borg, Walter R.; and Stone, David R., "Protocol Materials as a Tool for Changing Teacher Behavior," *Journal of Experimental Education,* 43:34-39, Fall 1974.
4. Borg, Walter R., "Protocol Materials as Related to Teacher Performance and Pupil Achievement" (in press).
5. Kounin, Jacob S., *Discipline and Group Management in Classrooms*, Holt, Rinehart and Winston, New York, 1970, 178 pp.

THE STABILITY OF THREE INDICES OF RELATIVE VARIABLE CONTRIBUTION IN DISCRIMINANT ANALYSIS

CARL J. HUBERTY
University of Georgia

ABSTRACT

An empirical comparison is made of three proposed indices of relative predictor variable contribution: (1) the scaled weights of the first discriminant function; (2) the total group estimates of the correlations between each predictor variable and the first function; and (3) the within-groups estimates of the correlations between each predictor variable and the first function. It was found that given a single run of an experiment, none of the indices was sufficiently reliable in identifying the rank-order of the variables except possibly when the total sample size was very large.

AS DESCRIBED BY Cooley and Lohnes (4, Chp. 9), multiple group discriminant analysis strategy begins with a principal axis analysis. This analysis is made, not of the predictor variable intercorrelation matrix, but of the matrix product $E^{-1}H$, where E and H are the $(p \times p)$ pooled within-groups and the between-groups deviation score cross-products matrices, respectively, and p is the number of predictor variables. This "factoring" may be construed as a partitioning of the discriminatory power of the set of predictor variables into uncorrelated components, called discriminant functions. The vectors obtained from the eigenanalysis of $E^{-1}H$ define a discriminant space such that when points representing the group are located within it, these points are separated from each other to a maximum degree.

Sample estimates of the weights of the ith discriminant function are determined by the $(p \times 1)$ eigenvector $\mathbf{b}_i$, associated with the eigenvalue λ_i, from the determinantal equation

$$|E^{-1}H - \lambda I| = 0.$$

The equation which leads to the weights is

$$(E^{-1}H - \lambda_i I)\,\mathbf{b}_i = \mathbf{0}.$$

which is obtained as a result of maximizing the ratio of the mean square between-groups to the mean square within-groups, the mean squares being based on the discriminant function values. The maximum number of discriminant functions necessary to represent the group differences is the smaller of p and $k - 1$, where k is the number of criterion groups.

To find the dimensionality of the so-called "discriminant space," either the eigenvalues are subjected to a significance test (10:372-373), or a subset of the non-zero eigenvalues that accounts for a large percent, say 80%, of the total discriminating power of the predictor variables may be chosen. By analyzing the k group centroids in the discriminant space, it is possible to determine the role of each of the discriminant functions retained. That is, some insight into the question, "Between what groups or sets of groups does each function discriminate?" may be gained, and it is often useful to determine which predictor variables are contributing the most and the least to such discriminations. The problem thus arises of defining suitable indices of predictor variable potency, in terms of relative variable contribution to discrimination.

One index that has been proposed by several writers, e.g., Tatsuoka (12), is based on the sample "beta" weights, that is, weights applicable to standardized predictor values. The standardized weights for the ith eigenvector are determined by multiplying each element of the original vector by the positive square root of the variance of the ith variable:

$$\mathbf{b}_i^* = aD\mathbf{b}_i \qquad [1]$$

where a is a scalar and D is the $(p \times p)$ diagonal matrix of the positive square roots of the principal diagonal elements

of E. (The a-value in Eq. [1] indicates that the eigenvectors are only unique up to a constant of proportionality.) It is argued that these weights may be used to assess the relative contribution of the predictors in determining the ith discriminant scores.

Another approach to the problem of assessing relative predictor variable contribution to discrimination involves estimates of the correlations between each of the predictors and each of the discriminant functions. Two estimates of these correlations have been used. When the data collected are considered representative of a single population, these "structure" correlations are based on the "total group" predictor intercorrelation matrix. The ith $(p \times 1)$ vector of these correlations is given by

$$\mathbf{r}_i = D_1 R D_2 \mathbf{b}_i \qquad [2]$$

where

D_1 = $(p \times p)$ scalar matrix of the reciprocal of the standard deviation of the scores on the ith discriminant function;
R = $(p \times p)$ "total" intercorrelation matrix of the p predictors;
D_2 = $(p \times p)$ diagonal matrix of "total" standard deviations of the p variables; and
$\mathbf{b}_i$ = $(p \times 1)$ vector of weights for the ith discriminant function.

Correlations computed this way are precisely the Pearson product-moment correlation coefficients between the sample predictor scores and the sample discriminant scores on the ith function (5:339).

If the underlying model is one of k populations with identical covariance matrices, then the maximum likelihood estimate of the true ith correlation vector is given by the $(p \times 1)$ vector [(2:53) or (9:225)],

$$\mathbf{r}_i^{\dagger} = D^{-1} (E\mathbf{b}_i) (\mathbf{b}_i' E \mathbf{b}_i)^{-1/2} \qquad [3]$$

where D is defined as in Eq. [1] and $\mathbf{b}_i$ is defined as in Eq. [2].

The purpose of this study was to investigate the stability, over repeated sampling, of three indices of relative predictor variable potency:

1. the scaled weights as given by Eq. [1];
2. the correlations as determined by Eq. [2]; and
3. the correlations as determined by Eq. [3].

Only the vector of scaled weights associated with the first discriminant function was considered in the present study. Reasons for this restriction are that the first function usually accounts for a large portion of the discriminatory power of the set of predictors, and that for each replication of the experiment (to be discussed in the next section), the number of "significant" functions may not be the same, although there will always be at least one. [See also Bargmann (1).]

Simulation Procedure

To effect the simulation of drawing random samples of size N from k p-variate normal populations with a known common covariance matrix, a high-speed electronic computer—IBM system 360, Model 65—was used. In this study the number of predictor variables considered was p=10, and the numbers of criterion groups were k = 3, and k = 5.

Standard normal scores were considered in the simulated sampling. In determining the common population covariance matrix Σ, the goal was to get covariances (i. e., correlations) that are typical of those found in applications. [Cochran (3) implies that in practice most correlations are positive and modest in size.] The classical factor analysis model (6:15) was considered in arriving at Σ:

$$\Sigma = A_{pop} A_{pop}' + D_{pop}^2 ,$$

where

A_{pop} = $(10 \times m)$ matrix of coefficients of m common factors (i. e., matrix of factor loadings), and

D_{pop} = (10×10) diagonal matrix of coefficients of the unique factors.

The communality of each of the predictors was arbitrarily set at .75, thus making the reliability of each predictor at least .75. This condition yields a D_{pop} matrix with all diagonal elements equal to .50.

Separation between the k populations was accomplished by prescribing a $(10 \times k)$ population weight matrix W_{pop}, and then obtaining the population mean matrix [see Rao (11:488)][1]:

$$M_{pop} = \Sigma W_{pop} \qquad [4]$$

(The use of the different Fisher and Rao discriminant analysis models is recognized—the relationship in Eq. [4] was merely used to get the desired separation. Even though the weights in the two models are not directly related, except in the two-group case, use of population counterparts of the Rao weights was considered appropriate in this study.)

Total sample sizes considered, across all k groups, were N = 90, 150, 300, and 450. Equal-group sample sizes $N_g = N/k$ were used. Corresponding to each N, sample score matrices of size $(10 \times N_g)$ were generated from each of the k p-variate normal populations having the common covariance matrix Σ. To generate these sample score matrices, a procedure similar to that suggested by Kaiser and Dickman

(7) was employed. A number was selected from a uniform (0, 1) distribution using a subroutine called RANDU, corresponding to which a number from a normal (0, 1) "continuous" distribution was located. This technique was used to produce the elements of both an $(m \times N_g)$ matrix $\hat{F}$ and a $(10 \times N_g)$ matrix $\hat{U}$. The subsample score matrix corresponding to group g was then obtained using

$$X_g = A\hat{F} + D\hat{U} + M_g$$

where

X_g = the $(10 \times N_g)$ matrix of "observed" scores, and

M_g = the $(10 \times N_g)$ matrix, the ith row of which contains the (constant) value of element (i, g) in M_{pop}

Thus, in essence, random samples were selected on the orthogonal F and U matrices, and the observed scores were obtained by the above transformation. This sampling experiment was repeated 100 times for each N-value to provide data for empirically checking both the reliability (in the sense of *consistency)* and the validity of the three indices under consideration.

Data Analysis

For each replication of the experiment the 10 predictor variables were rank-ordered according to the absolute value of each index. The criterion used to judge the stability of the three indices of relative predictor variable potency was the consistency of the observed rank of each variable over repeated replications of the experiment. A necessary but not sufficient essential for a valid index of variable potency is that it exhibits consistency over repeated sampling. That is, an index lacking such consistency provides no basis for inferential statements concerning the respective rank of selected predictor variables. Only the rankings of the variables with respect to the first discriminant function were determined.

These potency rankings were analyzed in two ways. First, the number of times each variable attained a given rank was determined. These number-of-times-per-rank counts were found for each of the four N-values for each of the two k-values studied. These counts were organized into 24 (3 indices $\times$ 4 values of $N \times$ 2 values of k) two-way contingency tables, the rows corresponding to the ten possible ranks and the columns corresponding to the ten variables.

The second analysis involved a correlational approach. For each index the ranks of the variables, with respect to the first discriminant function, were found for each replication of the experiment. Ranks from 1 to 10 were assigned according to the numerical value of the index. Thus, for each index, a 100-by-10 two-way table was formed for each value of N and each value of k. The relationship among the 100 rankings was determined by computing the coefficient of concordance W (8:95). This coefficient was computed for the first discriminant function, for each of the three indices, and for each of the four values of N in both a three- and five-group situation. This resulted in the computation of 24 coefficients in all. The significance of each observed value of W, i. e., the hypothesis that there was no consistency in the rankings over the 100 replications, was tested using a chi-square statistic (8:98). When an observed value of W was found to be significant, i. e., when there was evidence of some agreement of the potency rank-orderings of the discriminatory variables over repeated sampling, an estimate of the true ranking was obtained by ranking the variables according to the sums of the ranks allotted over the 100 replications. Kendall (8:114) has shown that this procedure gives a "best" estimate in a least-squares sense.

Results

Three Group Case (k = 3)

It was clear from tables exhibiting the number of times each variable attained a given rank for each index that the stability of the indices over repeated sampling is not very marked. If an index is operating consistently over repeated sampling, then each column of such a table would contain only one value which is large in relation to the others; such a pattern was not observed.

The W-values and the observed chi-square values corresponding to them are given in Table 1. All of the values were significantly different from zero (at the .01 level).

Table 1.—Coefficients of Concordance, W, and Associated Chi-Square Values for $k = 3$

Indices	W	χ^2_9
Index 1		
N = 90	.113	112.050
N = 150	.177	159.007
N = 300	.302	272.007
N = 450	.381	343.043
Index 2		
N = 90	.182	163.887
N = 150	.259	233.267
N = 300	.288	258.923
N = 450	.418	376.139
Index 3		
N = 90	.189	169.979
N = 150	.265	238.097
N = 300	.299	268.898
N = 450	.425	382.894

$\chi^2_9 = 21.666$ at 1% level.

Table 2.—Estimates of the True Rankings of the Predictor Variables for $k = 3$

Indices	Variable 1	2	3	4	5	6	7	8	9	10
Index 1										
N = 90	8	9	2	4	6	5	3	1	10	7
N = 150	6	7	2	5	8	4	3	1	10	9
N = 300	8	7	3	4	6	5	2	1	10	9
N = 450	8	7	2	4	6	5	3	1	10	9
Final	7.5	7.5	2	4	6	5	3	1	10	9
Index 2										
N = 90	10	2	7	3	8	6	4	1	9	5
N = 150	10	2	4	7	8	6	3	1	9	5
N = 300	10	2	5	4	8	7	3	1	9	6
N = 450	10	2	5	7	6	8	3	1	9	4
Final	10	2	5.5	5.5	8	7	3	1	9	4
Index 3										
N = 90	10	2	7	3	8	6	4	1	9	5
N = 150	10	2	4	7	8	6	3	1	9	5
N = 300	10	2	5	4	8	6	3	1	9	7
N = 450	10	2	5	7	6	8	3	1	9	4
Final	10	2	5	5	8	7	3	1	9	5

Some (when $N = 90$) of the observed W-values which, according to Kendall, may be interpreted as correlation (here in the sense of *reliability*) coefficients, are quite low. That these low values wre significant is simply a result of the power of the test to detect differences between the hypothetical zero-value of the population W-values and their observed values, which differences are of no practical consequence. It is clear that unless sample size is very large, neither the scaled weights nor either of the correlation estimates are very consistent over repeated sampling.

The population weights for Variables 9 and 10 were fixed at zero. That is, these two variables would be expected to exhibit minimum potency insofar as their contribution to discrimination among groups is concerned. Hence, it was possible to effect, to some extent, an evaluation of the validity of the three indices under consideration.

While none of the indices provides a very reliable rank-order of variable potency for a single run of the experiment, the reliability of each index is nevertheless sufficient to provide a reliable (in the sense of *consistent*) estimate of variable potency when the ranks are averaged over 100 replications of the experiment. Table 2 gives the potency of each variable based on the average value of its rank as assigned by each index over 100 runs of the experiment. With one exception, which occurred in the case of the smallest sample size ($N = 90$), Index 1 assigned potency ranks of least and next-to-least to Variables 9 and 10. Index 2, on the other hand, assigned potency ranks ranging from 4 to 6 to Variable 10, and the ranks assigned to this variable by Index 3 ranged from 4 to 7. Judged in the light of this criterion, Index 1 is clearly the most valid of the three.

As a check on the reliability of the average potency ranks over 100 replications of the experiment, Kendall's W was calculated for each index using the (average) ranks for the four sample sizes as given in Table 2. The W-values for the three indices were .95, .92, and .91, respectively. When the sum of the (average) ranks over the four sample sizes is used as a basis for assigning an overall potency rank to each variable, these "final" ranks are as shown in Table 2. On the basis of these final ranks, Index 1 again identified Variables 9 and 10 as least potent. However, one of these variables, 10, was assigned a rank of 4 by Index 2 and a rank of 5 by Index 3.

Five-Group Case (k = 5)

The results obtained in this case very closely parallel those obtained when the number of criterion groups was three. Values of Kendall's W were computed, as well as the chi-square values used in testing the significance of each. The results are reported in Table 3. The average potency ranks of each variable as asssigned by each index over 100 runs of the experiment are given in Table 4. Again, Index 1 assigned the lowest ranks to Variables 9 and 10. Index 2 and Index 3 performed somewhat better than in the three-group case, but they still failed to consistently identify Variable 9 as one of the two variables of lowest

Table 3.—Coefficients of Concordance, W, and Associated Chi-Square Values for $k = 5$

	W	χ^2_9
Index 1		
N = 90	.069	62.433
N = 150	.143	128.998
N = 300	.236	212.140
N = 450	.351	315.506
Index 2		
N = 90	.099	88.798
N = 150	.121	108.960
N = 300	.191	172.189
N = 450	.318	286.259
Index 3		
N = 90	.112	100.951
N = 150	.129	116.149
N = 300	.208	186.446
N = 450	.332	298.848

$\chi^2_9 = 21.666$ at 1% level.

potency. The value of Kendall's W for the (average) ranks as assigned by Index 1 over the four sample sizes was .85. The corresponding values for Index 2 and Index 3 were .95 and .97, respectively. Overall or "final" ranks were established, and are also given in Table 4. Again, Index 1 identified Variables 9 and 10 as least potent. Index 2 and Index 3 identified Variable 10 as least potent, but assigned final potency ranks of 7.5 and 8, respectively, to Variable 9.

It might have been well to relate the results obtained to the population (or true) character of the variables with regard to relative contribution. (It may be noted that in the true sense there is no agreed upon index of variable contribution.) No consistent relationships between the population weights (in W_{pop}) and the rank-orderings reported in Tables 2 and 4 resulted. In the three-group case, the best variable (Variable 8) had a large population weight for one group and two zero weights. The worst variable, as determined by function-variable correlations, was Variable 1 which had one large population weight and two moderate weights. Whereas, in the five-group case, the best variable (Variable 2), according to the correlation index, had all non-zero population weights. A low ranking variable (Variable 5) had two zero weights, two moderate weights, and one substantial weight.

Table 4.—Estimates of the True Rankings of the Predictor Variables for $k = 5$

Index	Variable									
	1	2	3	4	5	6	7	8	9	10
Index 1										
N = 90	3	6	8	4	5	7	2	1	10	9
N = 150	6	3	4	8	5	7	2	1	10	9
N = 300	4	1	5	8	7	6	3	2	9	10
N = 450	3	2	5	8	7	6	4	1	10	9
Final	4	3	5	8	6	7	2	1	10	9
Index 2										
N = 90	6	1	7	5	8	4	2	3	9	10
N = 150	8	1	6	5	9	4	2	3	7	10
N = 300	7	1	6	5	9	3	2	4	8	10
N = 450	8	1	6	7	9	3	2	4	5	10
Final	7.5	1	6	5	9	3.5	2	3.5	7.5	10
Index 3										
N = 90	5	1	7	6	8	4	2	3	9	10
N = 150	7	1	6	5	9	4	2	3	8	10
N = 300	7	1	6	5	9	3	2	4	8	10
N = 450	8	1	6	5	9	3	2	4	7	10
Final	7	1	6	5	9	3.5	2	3.5	8	10

Discussion and Conclusion

As in multiple regression analysis, the notion of variable contribution in discriminant analysis is an evasive one. In both analyses the variables act in concert and cannot logically be separated to determine how much each variable contributes to prediction or to a discriminant function. Thus, an index of *absolute* contribution is, considering the present state of knowledge, out of the question. An index of *relative* variable contribution is, however, an approach advanced by many researchers. Traditionally, the index used to gauge the contribution of each variable in the company of all others is the standardized discriminant weights of Eq. [1]. Some writers have proposed the correlations of Eq. [2] (4) and the correlations of Eq. [3] (2) for that purpose. Other writers, e.g., Tatsuoka (12), state that such correlations are not intended as measures of potency of discrimination, but as aids in "interpreting" resulting discriminant functions.

Both types of indices have been advanced for use as descriptive indicators of relative variable contribution. Some researchers have implied in discussing discriminant analyses, however, that their results can be expected to be found with other samples of subjects. That is, it is implied that the index of relative potency will rank-order the variables similarly across repeated sampling; further, it is assumed that if a variable is judged, on the basis of the descriptive index, to be a nondiscriminator, then a similar judgment would be made if data are collected on new subjects.

This study dealt with the reliability and, to some extent, the validity of the three proposed indices of relative variable contribution in discriminant analysis. Conclusions are limited to a situation in which (1) the k populations are 10-variate normal; (2) the k population covariance matrices are identical; (3) the number of "subjects" drawn from each population is the same; (4) only the first discriminant function is evaluated; and (5) elements of the common population covariance matrix and the differences between mean vectors are similar in magnitude to those used. In this situation the findings support the following conclusions:

1. Indices 2 and 3 can be expected to have very comparable reliability in assessing the relative potency of predictor variables; when the number of criterion groups is three, this reliability is slightly higher than that of Index 1, and vice versa for five criterion groups.

2. Index 1 can be expected to be the most valid in identifying those variables that contribute minimally to the discrimination involved.

3. Given a single run of the experiment, none of the indices can be expected to be sufficiently reliable to be of great practical value in identifying potent variables unless the total sample size is very large. The lack of reliability of the discriminant function-variable correlations found in this study contradicts a conclusion reached by Thorndike and Weiss (13), an investigation that involved two sets of real data.

FOOTNOTE

1. The A_{pop}, Σ, W_{pop}, and M_{pop} matrices are available from the author, Dr. Carl J. Huberty, College of Education, The University of Georgia, Athens, Georgia 30602.

REFERENCES

1. Bargmann, R. E., "Exploratory Techniques Involving Artificial Variables," in P. R. Krishnaiah (ed.), *Multivariate Analysis II,* Academic Press, New York, 1969, pp. 567-580.
2. Bargmann, R. E., "Interpretation and Use of a Generalized Discriminant Function," in R. C. Bose, et al. (eds.), *Essays in Probability and Statistics,* University of North Carolina Press, Chapel Hill, 1970.
3. Cochran, W. G., "On the Performance of the Linear Discriminant Function," *Technometrics,* 6:179-190, 1964.
4. Cooley, W. W.; and Lohnes, P. R., *Multivariate Data Analysis,* Wiley, New York, 1971.
5. Gulliksen, H., *Theory of Mental Tests,* Wiley, New York, 1950.
6. Harman, H. H., *Modern Factor Analysis,* University of Chicago Press, Chicago, 1967.
7. Kaiser, H. F.; and Dickman, K., "Sample and Population Score Matrices and Sample Correlation Matrices from an Arbitrary Population Correlation Matrix," *Psychometrika,* 27: 179-182, 1962.
8. Kendall, M. G., *Rank Correlation Methods* (3rd Ed.), Hafner, New York, 1962.
9. Porebski, O. R., "Discriminatory and Canonical Analysis of Technical College Data," *The British Journal of Mathematical and Statistical Psychology,* 19: 215-236, 1966.
10. Rao, C. R., *Advanced Statistical Methods in Biometric Research,* Wiley, New York, 1952.
11. Rao, C. R., *Linear Statistical Inference and Its Applications,* Wiley, New York, 1965.
12. Tatsuoka, M. M., "Multivariate Analysis in Educational Research," in F. N. Kerlinger (ed.), *Review of Research in Education,* Peacock, Itasca, Ill., 1973.
13. Thorndike, R. M.; and Weiss, D. J., "A Study of the Stability of Canonical Correlations and Canonical Components," *Educational and Psychological Measurement,* 33: 123-134, 1973.

COGNITIVE CONSISTENCY THEORY AND STUDENT EVALUATION OF TEACHER EFFECTIVENESS[1]

ROLPH E. ANDERSON
Drexel University

KWANG S. CHOI
Old Dominion University

JOSEPH F. HAIR, JR.
University of Mississippi

ABSTRACT

To determine the applicability of cognitive consistency theory to student evaluations of instructors, the grade expectations and instructor evaluations of 108 students were obtained during the first and last weeks of the term in five separate business management courses. Respondents were divided into three categories depending upon whether their grade expectations remained the same, fell, or rose from the beginning to the end of the course. Results supported cognitive consistency theory in that student evaluations of instructors varied directly, after disconfirmed grade expectations, with the directional change (up or down) in student grade expectations. The positive impact on instructor evaluations of an increase in student grade expectations was approximately the same as the negative impact of a decline in grade expectations.

ONE OF THE MOST important but controversial subjects in academia today is evaluation of faculty teaching effectiveness. Although instructor self-improvement is or should be the primary purpose of these evaluations, surpluses of faculty candidates and tight budgets have pressured administrators to seek some quantitative (seemingly objective) index by which faculty teaching effectiveness can be measured for decisions on individual retention, salary increases, promotion, and tenure. Probably, the most widespread method, since its introduction in the early 1900s, has been faculty rating questionnaires completed by students. But, just how valid are rating questionnaires? The objectives of this paper are to briefly critique faculty rating questionnaires, review some of the latest findings on their use, and present the results of an empirical study testing the theory of cognitive consistency as a predictor of student evaluations of college teachers.

The Problem

Faculty Rating Questionnaires

Many different types of rating forms have been utilized by colleges and universities, but frequently the ratings have provided little help to the instructor in improving his teaching skills. Questionnaires sometimes simply ask students to rate the instructor on some dimension (supposedly related to effective teaching), such as "accessibility for individual conferences," using a scale ranging from excellent to poor. But, compared to what? Many questionnaires ignore this critical scaling question, while others ask for comparisons with the "best," "average," or "worst" teachers. Such shifting or sliding reference points (which depend upon the student's personal sample of teachers) make comparison of student evaluations invalid. Use of the amorphous, unattainable concept of the "ideal" teacher as a reference point is probably little better since it too will vary widely among students. A few rating forms use a forced-choice format with specific descriptive phrases which serve as brief critiques of instructor performance and help reduce rater bias, especially the "halo" effect where the student rates the instructor alike on different teaching dimensions because of his overall attitude toward the instructor.

Oftentimes, it is not clear to the student raters just what is being measured (instructor characteristics, course content or material, methodology, changes in student knowledge, or personal objectives achieved). In addition, there is no reliable evidence that those qualities listed on a scale are the ones that advance or impede student attainment of specific educational objectives (5). Seldom is the amount of student learning taking place related to instructor effectiveness. To date, "teacher effectiveness" has not been adequately defined, and many stereotypes persist among students, faculty, and administrators about what makes an effective teacher.

Review of the research literature on teaching evaluation reveals few areas of solid ground. Descriptions of teaching effectiveness tend to vary according to who is doing the evaluation (4). For example, Wedeen (17) found that two sections of the same course taught concurrently by the same instructor with identical assignments and examinations can be perceived differently by different groups of students. In a recent study (1), it was revealed that student evaluations of instructors may even decline with increasing age differentials between students and instructors. In general, many rating forms have been found statistically reliable (i. e., repeatable), but the question of validity (i. e., what is being measured) remains unresolved (5).

Static vs. Dynamic Measurement

One problem with many of the studies on instructor evaluation has been the static as opposed to dynamic nature of their approaches. Albeit student expectations have been recognized as an important influence on the evaluation process (3, 8), the impact of "changing" student expectations has been largely ignored. Even the vital question regarding the effect of expected grades on student evaluations of instructor effectiveness has been dealt with largely from a static framework. Voeks and French (15) investigated the proposition of whether or not students are influenced by grades when they rate quality of teaching. Their finding was that grades and student ratings of instructors had no reliable relationship. More recently, Krull and Crooch (10:9) report that "most studies indicate that the relationship between the grade expected in a course and a student's evaluation of the instructor's teaching effectiveness show little positive correlation."

Indeed, it may well be that there is no significant relationship between the final grade a student receives, or expects to receive, and his evaluations of the instructor and course. That is, two individuals receiving or expecting final grades of C may rate the instructor quite differently; so might two students expecting final grades of A. Similarly, a student expecting a final grade of B may rate the instructor identically with someone expecting a final grade of D. The unanswered question in previous studies is: What is the effect of disconfirmed (or changed) grade expectations on student evaluations of instructors and courses? To illustrate, will there be a difference in the ratings by students who expected an A at the beginning of the course, but whose expectations have lowered to C near the end of the course? Conversely, what is the effect on ratings by a student who originally expected a C but whose expectations at the end of the course rose to A? It is this dynamic view of grade expectations which may be more valuable in accounting for student differences in perceptions of teacher effectiveness than the typical static approach.

Expectations

Expectations may be described as subjective notions of things to come (9). In terms of student relationships with instructors, an expectancy may be thought of as an initial hypothesis formed by the student, and his perception of the outcome after completing the course of instruction will serve to either confirm or reject the original hypothesis (6). Grade expectations are confirmed when the student receives the grade he originally expected. Negative disconfirmation results when the grade outcome is lower than prior student expectations. Positive disconfirmation occurs when the grade actually exceeds earlier expectations.

When expectations are realized, student evaluations of the instructor and course should coincide with prior expectations and ratings, but what are the effects on evaluations when the student's expectations for a grade have been disconfirmed, either negatively or positively?

Disconfirmed Expectancies and the Theory of Cognitive Consistency

In determining the impact of disconfirmed expectations, the psychological theory of cognitive consistency deserves

major consideration. Stated briefly, cognition refers to the constellation of knowledges, beliefs, attitudes, and perceptions that an individual has concerning himself, his behavior, and his milieus (13). From the individual's cognitions, a conceptual framework is developed which guides his behavior (11). At the heart of all cognition-related theories of human behavior is the premise that individuals seek consistency among cognitions. That is, they strive for compatibility among their knowledges, beliefs, values, and perceptions of themselves, and elements or persons in their physical or psychological frame of reference. Inconsistent cognitions create psychological stress or tension which, for relief, compels behavior directed toward the attainment of consistency. Osgood and Tannenbaum's congruity principle (12) explains this process more precisely. It implies that when two or more objects (a communication source and a goal object) are associated or linked together by an assertion, there is a tendency for the evaluation of one or both objects to change so that the two evaluations become more alike. The principle further stresses that changes in evaluation will always occur in the direction of increased congruence within the individual's frame of reference.

Students continuously receive various kinds of feedback from their own experiences, peers, instructors, and classroom performances. These information inputs are cognitions which students like to keep consistent with one another. When a student receives two pieces of information which are psychologically dissonant, he attempts to reduce this mental discomfort by changing or distorting one or both of the cognitions to make them more consonant or compatible. The more powerful the cognitive dissonance, the more inclined he is to attempt to reduce dissonance by changing the cognitive elements (2). Accepting the premise of cognitive consistency, pre-course expectations of students regarding their individual performance and the quality of instruction would tend to coincide since the students are free to adjust either expectation to achieve consonance. However, any discrepancy between student grade expectations and ratings at the beginning of the course compared to the grade expectations at the end of the course will likely be resolved by the student's adjusting his perceptions of the course and the instructor so that evaluations become more consistent (less dissonant) with his final expectations.

To test the applicability of cognitive consistency theory to the teacher evaluation process, the following hypotheses were considered when student grade expectations are disconfirmed:

1. *Null Hypothesis:* Instructor and course evaluations by students before and after disconfirmed grade expectations are not significantly different.

2. *Research Hypothesis:* Instructor and course evaluations by students tend to vary directly, after disconfirmed grade expectations, with the directional change (up or down) in student grade expectations.

Method

To test the hypotheses, 140 undergraduate and graduate business students were asked to record their expectations on an instructor and course evaluation questionnaire at the start of five separate courses in marketing and finance. Then, during the last week of classes, the students recorded their final expected grades and instructor/course evaluations on the same questionnaires, illustrated in Table 1. Complete student anonymity was assured by allowing students to select their own codes for the two questionnaires for later matching of the initial and final ratings. For several reasons (dropouts, unmatched questionnaire codes, or absenteeism on the day of evaluations), the final study was reduced to 108 students.

The students were divided into three groups: (1) those whose grade expectations remained the same in both surveys; (2) those whose expectations fell from the first to the second survey; and (3) those whose grade expectations rose. This taxonomy placed 61 students in the "no-change" group, 33 in the "downward-change" category, and 14 in the "upward-change" group. Mean responses by students in each of the three grade expectations categories were compared on the sixteen questionnaire variables by univariate F-tests to determine significant differences.[2] For decision purposes in this study, a significance level of .10 or higher was chosen.

Results

To determine what effect, if any, disconfirmed grade expectations may have on student ratings of instructors and courses, the first and second evaluations by students in each separate group were compared.

No-Change Category

As shown in Table 2, there was a significant difference (.05 level or beyond) in ratings between the first and second evaluations on eight (Nos. 1, 2, 5, 7, 10, 11, 12, 16) of the sixteen variables concerning student perceptions of the instructor and course. Four of the rating variables rose and four fell from the first to the second evaluations. The t-test comparing the composite mean scores of the first and second evaluations yielded a value of .7135, which is not significant. Thus, it appears that there is no systematic change pattern in student perceptions when grade expectations remain constant during a course of instruction.

Downward-Change Category

Nine rating variables (Nos. 1, 2, 4, 7, 8, 9, 12, 13, 14), as can be seen in Table 3, were significantly different between the first and second surveys in the group whose grade expectations changed downward during the term. All nine of these variables fell, i. e., they moved in the direction of the change in student grade expectations. It is also interesting to note that 15 of the 16 variables changed downward in the direction of the negatively disconfirmed grade

Table 1.–Instructor and Course Evaluation Forms

VARIABLES	Always	Usually	Sometimes	Seldom	Never	Cannot Comment
THE INSTRUCTOR:						
1) Was well prepared for class	5	4	3	2	1	0
2) Stimulated intellectual curiosity	5	4	3	2	1	0
3) Encouraged independent chinking	5	4	3	2	1	0
4) Was concerned that the class understood him	5	4	3	2	1	0
5) Showed respect for the questions and opinions of students	5	4	3	2	1	0
6) Was accessible for individual conferences	5	4	3	2	1	0
7) Did the tests cover what you could reasonably be expected to know?	5	4	3	2	1	0
8) Were the lectures relevant to the course?	5	4	3	2	1	0
9) Were required assignments appropriate to the course?	5	4	3	2	1	0
10) Was discussion used or allowed when appropriate to the course?	5	4	3	2	1	0
11) Did the instructor have a pleasant personality?	5	4	3	2	1	0
12) Rate this course as a beneficial educational experience	5	4	3	2	1	0
13) Rate the quality of the instructor's presentation	5	4	3	2	1	0
14) Rate the overall teaching ability of this instructor	5	4	3	2	1	0
15) How was the instructor of this course recommended by other students?	5	4	3	2	1	0
16) How fair (objective) do you think the instructor of this course will be in grading you?	5	4	3	2	1	0

Table 2.–No Change in Grade Expectations Mean Ratings

Variables	First Evaluation Means	Second Evaluation Means
1. PREPARATION	4.5738	4.2295[a] ↓
2. STIMULATION	4.3279	3.9836[a] ↓
3. THINKING	4.2787	4.1967
4. COMMUNICATED	4.5574	4.5082
5. RESPECT	4.4918	4.7377[a] ↑
6. ACCESSIBLE	4.1803	4.0984
7. TESTS	4.4754	4.0164[a] ↓
8. ASSIGNMENTS	4.4754	4.3115
9. LECTURES	4.4426	4.2295
10. DISCUSSION	4.4098	4.6230[b] ↑
11. PERSONALITY	4.3115	4.8361[a] ↑
12. EDUCATIONAL	4.3279	3.8361[a] ↓
13. PRESENTATION	4.2295	4.0000
14. OVERALL	4.2459	4.1475
15. RECOMMENDED	3.0984	2.7213
16. FAIRNESS	4.2131	4.4754[b] ↑
GRAND MEANS	4.2900	4.1844
STANDARD DEVIATIONS	.3303	.4682

[a] $p < .01$ [b] $p < .05$

Table 3.–Downward Change in Grade Expectations Mean Ratings

Variables	First Evaluation Means	Second Evaluation Means
1. PREPARATION	4.5455	4.2121[b] ↓
2. STIMULATION	4.2424	3.3333[a] ↓
3. THINKING	4 .515	3.8788
4. COMMUNICATED	4.5152	4.0606[a] ↓
5. RESPECT	4.6667	4.5152
6. ACCESSIBLE	4.4545	4.1515
7. TESTS	4.6364	3.8788[a] ↓
8. ASSIGNMENTS	4.6970	4.3333[c] ↓
9. LECTURES	4.6061	4.2727[b] ↓
10. DISCUSSION	4.3636	4.2727
11. PERSONALITY	4.3333	4.5152
12. EDUCATIONAL	4.3939	3.3030[a] ↓
13. PRESENTATION	4.3030	3.5758[a] ↓
14. OVERALL	4.2727	3.6667[a] ↓
15. RECOMMENDED	2.2727	2.0909
16. FAIRNESS	4.3333	4.0303
GRAND MEANS	4.2992	3.8807
STANDARD DEVIATIONS	.5466	.5863

[a] $p < .01$ [b] $p < .05$ [c] $p < .10$

expectations. This mass movement of student ratings on the individual questionnaire items suggests the presence of a negative "halo" effect consistent with the students' lowered grade expectations. The *t*-test score of 2.023 with 30 degrees of freedom is significant at the .10 level for the two-tailed test and just short of the 2.042 needed for significance at the .05 level of significance.

Upward-Change Category

In Table 4, only four of the rating variables (Nos. 5, 8, 10, 15) were significantly different between the first and second evaluations. But, again, all four of the variables changed upward in the direction of the positively disconfirmed grade expectations of the students. This time it seems a positive "halo" effect was operating, as 14 of the 16 rating variables changed upward (or remained the same in two cases) in the direction of the revised student grade expectations. Only two of the sixteen variables moved in the opposite direction. The *t*-value of 3.299 with 30 degrees of freedom is significant at the .001 level.

Regression Analysis with Binary Variables

In order to obtain additional information about the precise nature of the relationship between student grade expectations and instructor evaluations, regression analysis was employed with binary variables as independent variables. When using binary variables exclusively as independent variables, it can be shown that regression analysis amounts to a variation of an analysis of variance test.[3]

The dependent variable used in the model represented the difference in mean evaluation scores between the first and second surveys for each of the 16 questions, for each of the three groups. This is denoted as $d_i = y_{2i} - y_{1i}$, where $i = 1 \ldots, 48$. There are thus a total of 48 observations available for use with the regression equation. The first independent variable, X_1, is defined to be a 1 (and 0 otherwise) for each of the 16 values of d_i associated with the group of students whose grade expectations rose. The second independent variable, X_2, is defined to be a 1 (and 0 otherwise) for each of the 16 values of d_i associated with the group of students whose grade expectations fell over the semester. It can be shown that the effect of defining the regression in this way is to produce the following result:

$$\hat{d} = \mu_{NC} + (\mu_{EF} - \mu_{NC}) X_1 + (\mu_{ER} - \mu_{NC}) X_2$$

where:

$\hat{d}$ = the calculated difference in evaluation scores between the first and second surveys

μ_{NC} = the average value of d_i for the "no-change" group

μ_{EF} = the average value of d_i for the group whose expectations fell

Table 4.—Upward Change in Grade Expectations Mean Ratings

Variables	First Evaluation	Second Evaluation
	Means	Means
1. PREPARATION	4.4286	4.2857
2. STIMULATION	4.2857	4.2143
3. THINKING	4.1429	4.5000
4. COMMUNICATED	4.0000	4.3571
5. RESPECT	4.2143	4.6429[c] †
6. ACCESSIBLE	3.9286	4.3571
7. TESTS	4.1429	4.2143
8. ASSIGNMENTS	4.0714	4.5000[c] †
9. LECTURES	4.2143	4.4286
10. DISCUSSION	4.1429	4.6429[b] †
11. PERSONALITY	4.3571	4.5714
12. EDUCATIONAL	4.1428	4.1428
13. PRESENTATION	4.0714	4.1428
14. OVERALL	3.9285	4.2857
15. RECOMMENDED	3.6667	4.0000[a] †
16. FAIRNESS	4.3571	4.3571
GRAND MEANS	4.1310	4.3527
STANDARD DEVIATIONS	.1855	.1828

[a] $p < .01$ [b] $p < .05$ [c] $p < .10$

μ_{ER} = the average value of d_i for the group whose expectations rose[4]

The constant term in this model represents the average value of d_i for the "no-change" group, while the coefficients of X_1 and X_2 represent, respectively, the difference between μ_{EF} and μ_{NC}, and μ_{ER} and μ_{NC}.

Table 5 summarizes the results of the experiment. The average value of d_i (−.106) for the "no-change" group is not significantly different from zero, which means that student evaluation scores did not fall significantly between the first and second surveys among those students whose grade expectations did not change. However, the coefficients for X_1 and X_2 are significantly different from zero (beyond the .01 level), indicating that the difference between μ_{EF} and μ_{NC} and the difference between μ_{ER} and μ_{NC} are both significantly different from zero. What this means is that the average difference between the first and second evaluation scores was significantly lower, relative to the average difference for the "no-change" group, for those whose grade expectations fell over the semester. Similarly, the second coefficient shows that the average difference between the first and second evaluation scores was significantly higher, relative to the average difference for the "no-change" group, for those whose grade expectations rose during the semester. One significant

finding that stands out is that the impact on the average difference in evaluation scores of a downward change in grade expectations appears to be about the same order of magnitude as the impact caused by an upward adjustment in grade expectations, although the effects are reversed. That is, the coefficients for both variables are virtually the same.

The explained sum of squares for this regression equation is simply the sum of squared deviations of group means from the overall mean of d_i. Thus, the R^2 for this regression will be large when between-group variation of d_i is large relative to the within-group variation. The value of R^2 for this regression indicates that approximately 50% of the variation in the dependent variable is being explained by the independent variables, X_1 and X_2.

Summarizing, during the semester, 61 of 108 sample students did not change their grade expectations, 33 students revised downward, and 14 adjusted their expectations upward. In the "no-change" group, the overall ratings of instructor and course did not significantly change from the first week to the last week of classes. The "adjusted-down" students significantly lowered (.01) their final overall ratings compared to initial ratings. Conversely, in the "adjusted-up" group, students significantly raised (.001) their evaluation on the final survey. Initial ratings of instructors and courses were the lowest in the "adjusted-up" group, followed by the "no-change," and "adjusted-down" group. On the final evaluations, however, ratings were highest in the "adjusted-up" group, followed by the "no change" and "adjusted-down" groups. The absolute value of the adjustment score was the largest in the "adjusted-down" group, followed by the "up" and "no-change" groups.

Results were consistent with a hypothesis that a conservative initial (grade) expectation contradicted or disconfirmed by subsequent positive feedbacks tends to have a positive upward effect on instructor and course ratings by students. On the other hand, elevated initial (grade) expectations disconfirmed by subsequent negative feedback tend to negatively influence instructor and course evaluations. Interestingly, the positive impact on teacher and course evaluations of an increase in student grade expectations is approximately the same as the negative impact of a decline in grade expectations.

Table 5.–Calculated Relative Differences in Instructor Evaluation Scores for Student Grade Expectations Groups

	NC	EF	ER
μ	-.106	-.313	+.327
s	+.138	.098	.141
N	61	33	14
t	1.52	3.18*	3.32*

*$p < .01$

$\hat{d} = -.106 - .313X_1 + .327_2$
(.070) (.098) (.098) *standard error of the distribution of coefficients*

$R^2 = .48$

Discussion and Implications

All thirteen significantly different rating variables between the two surveys in both the downward-and upward-change student groups moved in the direction of the change in student grade expectations. Therefore, results of this empirical study strongly support the research hypothesis that instructor and course evaluations by students tend to vary directly, when grade expectations are disconfirmed, with the directional change (either up or down) in student grade expectations.

Several implications can be derived from this research for college and university teachers, administrators, and students alike. First, instructors would do well to guard against creating unrealistically high grade expectations for students at the beginning of a course. In fact, it may be that an opposite approach, i. e., generating low grade expectations initially, will lead to higher student evaluations of the instructor and course.[5] It is significant that more than twice as many of the rating variables (nine vs. four) fell as rose when student grade expectations were disconfirmed, either negatively or positively. Apparently, a downward decline in student grade expectations can be especially adverse to student evaluations of the instructor and course.

Obviously, much more needs to be discovered about student expectations and their effect on perceptions and evaluations. It is hoped that this study will stimulate further investigation of the impact of changing expectations on student evaluations of instructor performance.

Since instructor self-improvement should be the main purpose of student evaluations, with administrative decisions on faculty salary, promotion, and tenure secondary considerations, administrators should carefully review all existing evaluation systems to ensure that they are providing faculty with sufficient informational feedback to encourage action to improve teaching effectiveness. Mere rankings or percentile ratings on various dimensions provide little guidance to the instructor in determining ways to improve his performance.

Students should be made aware of their responsibilities by administrators and faculty. The students should be encouraged to exercise honest, mature judgment in evaluating instructors and courses as precisely and comprehensively as possible. Too often, the student views the evaluation process as an imposition on his time–a task he wants to complete as quickly as possible. Seldom is more than ten minutes allotted to this individual evaluation procedure,

yet the outcome can be critical to faculty careers, administrative decisions, and student satisfaction. . .and over the long-run, community progress.

Administrators must use caution and judgment in attempting to compare one instructor with another on the basis of student evaluations. To date, the available research findings on ratings of teaching effectiveness are "incomplete, inconclusive, and of limited value" (5). Much more reliability than validity exists in most present instructor evaluation systems. Further research on teaching effectiveness needs to be emphasized in colleges and universities in all departments, and findings disseminated to faculty, administrators, and students. This research should be conducted as part of a systematic and on-going program designed to identify relevant behavioral, psychological, or environmental variables directly related to student gains or objectives. Research may even indicate that a number of rating scales will be needed for specific uses, e. g., in specific subject areas, for different learning objectives, and for students of various backgrounds, levels, or need orientations.

Nearly all participants in the educational process agree that instructor evaluation is a necessary and healthy activity, but the procedures and techniques in obtaining and interpreting data from the evaluation process need to be improved. Working together, faculty, administrators, and students can create a cooperative, as opposed to a competitive, atmosphere where progress can be rapidly made toward the advantage of all participants.

FOOTNOTES

1. The authors wish to acknowledge the advice and assistance of Dr. Kenneth E. Galchus, Assistant Professor of Quantitative Sciences in Business and Economics at Old Dominion University, Norfolk, Virginia.

2. Program DSCRIM, see (14).

3. See Jan Kmemta, *Elements of Econometrics*, Macmillan, New York, 1971, 409-430.

4. *Ibid.*, pp. 410-415.

5. Even though the present study included only one rating variable–#12–dealing directly with the course itself rather than the instructor, it has been found that students tend to rate the course and teacher the same (16).

REFERENCES

1. Adams, H. L., "Favorable Student Evaluations as a Function of Instructor's Age," *Improving College and University Teaching*, 21:72, Winter 1973.
2. Brehm, J. W.; and Cohen, A. R., *Explorations in Cognitive Dissonance*, Wiley, New York, 1962.
3. Carpenter, F.; Egmond, E. V.; and Jochem, J., "Student Preference of Instructor Types as a Function of Subject Matter," *Science Education*, 49:235-238, 1965.
4. Crawford, P. L.; and Bradshaw, H. L., "Perception of Characteristics of Effective University Teachers: A Scaling Analysis," *Educational and Psychological Measurement*, 28:1079-1085, 1968.
5. Dwyer, F. M., "Selected Criteria for Evaluating Teacher Effectiveness," *Improving College and University Teaching*, 21:51-52, Winter 1973.
6. Engel, J. F.; Kollat, D. T.; and Blackwell, R. D., *Consumer Behavior*, Holt, Rinehart and Winston, New York, 1968, p. 512.
7. Festinger, L., *A Theory of Cognitive Dissonance*, Harper & Row, New York, 1957.
8. Jandt, F. E., "A New Method of Student Evaluation of Teaching," *Improving College and University Teaching*, 21:15-16, Winter 1973.
9. Katona, G., "Business Expectations in the Framework of Psychological Economics (Toward a Theory of Expectations)," in M. J. Bowman (ed.), *Expectations, Uncertainty, and Business Behavior*, Social Science Research Council, New York, 1958, p. 59.
10. Krull, G. W., Jr.; and Crooch, G. M., "Measuring Teaching Effectiveness: An Improved Instrument," *Collegiate News and Views*, 27:9-13, Fall 1973.
11. Markin, R. J., Jr., *Consumer Behavior*, Macmillan, New York, 1974, p. 130.
12. Osgood, C. E.; and Tannenbaum, P. H., "The principle of Congruity in the Prediction of Attitude Change," *Psychological Review*, 62:42-55, 1955.
13. Oshikawa, S., "The Theory of Cognitive Dissonance and Experimental Research," *Journal of Marketing Research*, 5:429-430, November 1968.
14. Veldman, D. J., *FORTRAN Programming for the Behavioral Sciences*, Holt, Rinehart and Winston, New York, 1967.
15. Voeks, V. W.; and French, G. W., "Are Student Ratings of Teachers Affected By Grades?", *Journal of Higher Education*, 31:330-334, 1960.
16. Weaver, C. H., "Instructor Rating by College Students," *Journal of Educational Psychology*, 51:21-35, 1969.
17. Wedeen, S. U., "Comparison of Student Reaction to Similar, Concurrent Teacher-Content Instruction," *Journal of Educational Research*, 56:540-543, 1963.

DIRECTIONS FOR J.E.E. CONTRIBUTORS

The Journal of Experimental Education publishes specialized or technical education studies, treatises about the mathematics or methodology of behavioral research, and monographs of major current research interest.

ABSTRACT REQUIRED

An abstract of not more than 120 words must accompany each manuscript. It should precede the text, be designated *ABSTRACT*, and conform to the following standards:

1. In a research paper, include statements of (a) the problem, (b) the method, (c) the data, and (d) the conclusions.

2. In a review or discussion article, state the topics covered and the central thesis.

3. Standard abbreviations may be used freely if the meaning is clear. Use complete sentences and do not repeat information which is in the title.

TEXT

Usually research articles should have a clearly organized order of presentation. The following elements and their order is a guide and is not intended to be prescriptive.

The Problem. The nature, scope, and significance of the problem should be presented.

Related Research. Presentation and discussion of related research should be selective and should include references essential to clarify the problem under examination.

Methodology. This section should consist of hypotheses, description of the sample and sampling procedures, discussion of the variables, description of the data-gathering instruments (if well-known, their description may be omitted), and general description of the statistical procedures.

Presentation and Analysis of Data. Analysis of the data and conclusions about the hypotheses should be more than mere presentation. Rival hypotheses and significance for related studies and educational theory should be considered. Summary data (means, standard deviations, frequencies, correlation coefficients, etc.) should be presented in tabular or graphic form. Discussion of these data should be interpretive.

Summarizing Statements. A summary of conclusions and implications for education may supplement the abstract.

STYLE

Certain suggestions are offered here to achieve uniformity in publication style and to conserve the time and energy of the author and editorial staff in the preparation of copy. *A Manual of Style,* 12th ed., University of Chicago Press, Chicago, 1960, may be used as a style manual in preparation of manuscripts.

Two Copies Required. Copies should be double spaced with wide margins. Typed copies are preferred, but dittoed or mimeographed copies will be accepted if they are legible.

Subheads. Articles are frequently improved by the judicious use of subheads. However, avoid use of the title, INTRODUCTION, for a lead section.

Title. Try to use a short title, preferably no more than ten words. Avoid superfluous phrases, such as "A Comparison of . . .," "A Study of . . .," and "The Effectiveness of . . . ."

Tables. Prepare tables precisely as they are to appear in *The Journal,* caption each with a brief and to-the-point title, and number consecutively with arabic numerals: *Table 3.* INTERCORRELATION MATRIX, VARIABLES OPTIMALLY ORDERED.

Figures. Draw any graphs and charts in black ink on good quality paper, title each, and number consecutively, thus: *Figure 4.* SCHOOL ENROLLMENT. Send the original copy. We cannot accept mimeographed figures or charts. Data should not be framed, and ordinates and abscissas should be shortened to occupy as little space as possible.

Tables and Figures. Tables and figures must be original copies acceptable for reproduction. A charge will be assessed for any redrawing or re-typing of tables or figures.

Technical Symbols. All symbols, equations, and formulas must be clearly represented. Differentiate between numbers and letters (zero and the letter o; one and the letter l, etc.) Identify hand-drawn Greek letters in the margin. Align subscripts carefully.

Footnotes. Avoid explanatory footnotes by incorporating their content in the text. However, for essential footnotes, identify them with consecutive superscripts, as *book,*[2] *study,*[3] etc., and list the footnotes in a section, entitled FOOTNOTES, at the end of the text, but preceding the REFERENCES.

References. References should be listed alphabetically according to the author's last name at the end of the manuscript. Each entry should be numbered. Citation of a reference in the text should be made with this number, as (2); or if specific pages are to be indicated, as (2:245–7).

Illustrations of article and book references:

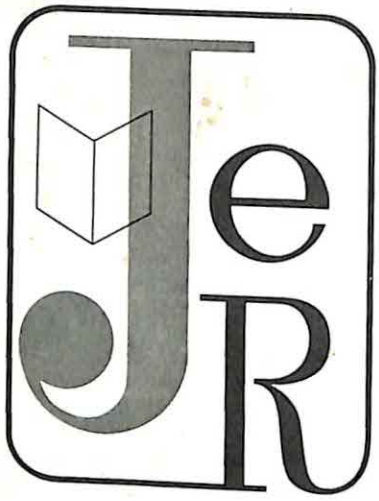
1. Bittner, Reign H. and Wilder, Carlton E., "Expectancy Tables: A Method of Interpreting Correlation Coefficients," *The Journal of Experimental Education,* 14:245–52, March 1946.
2. Thomson, Godfrey, H., *The Factorial Analysis of Human Ability,* Houghton Mifflin Co., Boston, 1950, 383 pp.

PROCEDURES

Send manuscripts to John Schmid, Department of Research and Statistical Methodology, University of Northern Colorado, Greeley, CO 80631.

Each contributor will receive 2 complimentary copies of the issue in which his article appears. Keep an exact carbon copy of your manuscript so that if a question arises the editors can refer you to specific pages, paragraphs, or lines for clarification by letter.

The Journal of Educational Research



THE JOURNAL OF EDUCATIONAL RESEARCH lives at the growing edge of a productive scholarship in the universe of education devoted to the discovery, documentation, and dissemination of the knowledge and insights by which new truths are found.

The journal presents early evidence on all the major breakthroughs in education: individual differences, learning and problems, training techniques, tests and measurements, curriculum, counseling and guidance, methods of teaching in all subjects, supervision, investigations of reading, teacher education, evaluating teacher effectiveness, administration, and other areas.

In scope and in stature, THE JOURNAL OF EDUCATIONAL RESEARCH has sought constantly to improve its services to its field. Over the past 50 years it has pioneered new formats that have made possible the more rapid and economical reproduction of more learned papers than had been thought possible.

Monthly with combined issues May/June and July/August. One year $15.00. Two years $30.00. Add $3.00 per year for subscriptions outside the United States and Canada.

THE JOURNAL OF EDUCATIONAL RESEARCH
Suite 302
4000 Albemarle Street, N.W.
Washington, D.C. 20016

Name _____

Street _____

City _____

State _____ Zip _____

THE JOURNAL OF EXPERIMENTAL EDUCATION

4000 Albemarle Street, N.W., Suite 302,
Washington, D.C. 20016

Return Postage Guaranteed

Second Class
Postage Paid at
Washington, D.C.

THE *Journal* OF

Experimental Education

Volume 44, Number 3 Spring 1976

In this issue:

THE JOURNAL OF EXPERIMENTAL EDUCATION

THE *Journal* OF EXPERIMENTAL EDUCATION

Volume 44, Number 3 Spring 1976

CONTENTS

The Journal of Experimental Education is published four times a year by HELDREF publications, 4000 Albemarle St., N.W., Washington, D.C. 20016. Annual subscription rates are $12.50 for institutions and $10 for individuals, plus $3 postage for all subscriptions outside the United States and Canada. Single copies $3. Second class postage paid at Washington, D.C. All business correspondence should be sent to this address. Claims concerning missing issues made within 6 months will be serviced free of charge. Send all manuscripts to Prof. John Schmid, Department of Research and Statistical Methodology, University of Northern Colorado, Greeley, CO 80631.

Arvil S. Barr, Founder

EDITOR AND PUBLISHER © 1932-1962

(The Journal of Experimental Education is indexed/abstracted in Abstr. S.W., CSPA, Current Contents, Ed. Adm. Abst., Educ. Ind., Soc. of Ed. Abst., Current Index to Journals in Education, Language and Language Behavior Abst.)

ENVIRONMENTAL NUMBNESS IN THE CLASSROOM[1]

ROBERT GIFFORD
University of Guelph
Guelph, Ontario

ABSTRACT

Following research on deleterious effects of surroundings on the behavior of users of other institutions, a naturalistic study of classroom-student interaction was conducted. Instructor-experimenters observed and recorded the behavior of university students in a laboratory which had been slightly altered to maximize difficulty of movement in the room. The amount and frequency of student alteration of the inhospitable furnishings was compared with person-furnishing distances in a non-institutional, personalized setting. The results indicated a strong tendency for students to accept without alteration a rather uncomfortable classroom arrangement. A brief discussion of possible implications for student attitudes toward school follows, one of these being that a specific inhospitality may lead to a diffuse negative feeling and may affect communicative behavior.

HOMO SAPIENS is commonly and correctly regarded as the greatest living shaper of the natural environment. Especially in contemporary society, when one gazes over a skyline, expressway, or dam, the image of "conqueror of the environment" (3) seems a truism. People have tremendously altered certain of earth's elements in active pursuit of the comforts and amenities first envisioned by the inventive members of the species.

Only recently has a broad awareness come that the shaper is, in part, shaped by his creations. When planners are proportionately few and users are not consulted in the design of a building, a danger arises that the needs of those who must spend large parts of their day in the structure will not be met. Mankind conquers while man lives; beneath the imposing skyline, micro-environments of dubious comfort and dignity exist. Not all these are in slums and ghettos; new and architectural award-winning structures have come in for their just share of the criticism (1).

Sommer (4) has discussed numerous types of buildings, including schools, in which the arrangement of furniture has apparently played an important role in behavior and communication. He found no malice on the part of institutional authorities, but rather an ignorance of the principles of design coupled with a *de facto* default in this matter on the part of maintenance workers, who are not interested in facilitation of learning in the users.

Users of public and semi-public buildings seem to develop an "environmental numbness" (5) to unpleasant sounds, sights, and arrangements. In one informal experiment, visitors were seated in front of a very annoying fan. None complained, but when finally the sound was consciously brought to their attention, nearly all acknowledged its unpleasantness. Sommer (4) feels that prolonged exposure to an institutional setting tends to lead to "institutional sanctity" or the feeling on the part of the user that whatever the setting, unpleasant or not, any change is regarded as improper by the user. The genesis of institutional sanctity is related to the same societal more which leads to the stiff suppression of user-initiated changes in the design of the environment, as in the case of People's Park in Berkeley and the custodian's everyday war on graffiti.

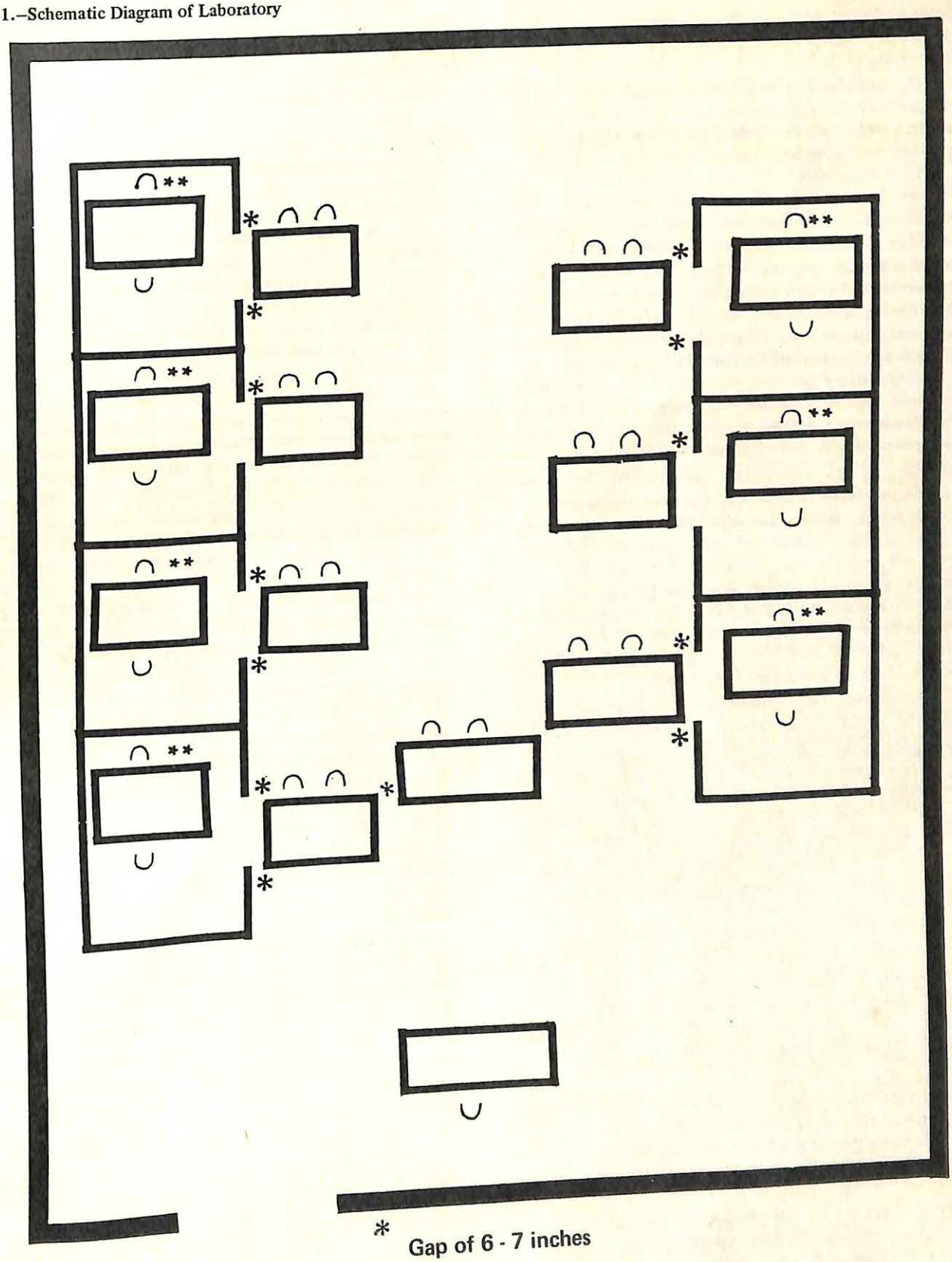
A previous study (2) of student reaction to surroundings found that students will, in their dormitories, handle an unliked but school-owned piece of furniture (a stiff chair and study desk) by ignoring it. Most students, and especially those with the higher grade-point averages, were found to study on the floor, the bed, or a lounge chair. The present study was designed to investigate student behavior in a situation where no alternative was available; they had to either accept or alter their surroundings. It was hypothesized that in the institutional setting, the students would not substantially change an inhospitable arrangement, despite their membership in a relatively high environmental awareness group.

Method

A university course in experimental psychology routinely conducts a didactic experiment in short-term memory early in each semester. The laboratory used for this purpose affords an opportunity for naturalistic observation and appropriate mis-arrangement of furnishings. The tables and chairs in the laboratory are of light construction, which no student would have difficulty in moving. The floor consists of waxed and polished tile.

The laboratory procedure calls for a pre-experimental discussion, the memory experiment, and a post-experiment discussion of results. The first and third parts take place in the open center area of the room (see Figure 1), and the

Figure 1.–Schematic Diagram of Laboratory



* Gap of 6 - 7 inches

** Seating Space 12 inches

experiment occurs in cubicles constructed from movable partitions along the sides of the room. During the experiment, pairs of students test each other, alternating *E* and *S* roles.

Prior to four laboratories sections, the tables and cubicles were carefully arranged as in Figure 1. This arrangement was on the whole the same as the students had previously experienced, except that all passages were reduced to the minimum necessary for a slender to medium-sized male to pass through without being forced to move the furniture. Thus, at each asterisked gap in Figure 1, the distance between articles of furniture was set at 6-7 inches. In order to pass, as the students had to do to move from central tables to cubicles, it was necessary to turn sideways and, depending on girth, maneuver carefully to avoid moving furniture. Of course, all students also had to pass through the frontal barrier from the entryway to their seating places at the beginning and end of the class, or to consult with the teacher if he happened to be in front of the barrier of tables.

Inside the cubicles, further "tight squeezes" were present. One side of each cubicle was arranged so that the distance between the back of the chair and the nearest table edge was 12 inches. The other chair, because of the size of the cubicles, had a much greater distance, allowing its occupant plenty of room to move. Although this 12-inch distance seemed certainly too close for ease of movement and comfort, a small naturalistic investigation of chair-table distances was made in situations where people felt free to set this distance, or at least had no impediment physically blocking the back of their chair.

In an academic office complex, an observer walked through the corridors and at every open door quietly asked the occupant, if they were involved in deskwork, not to move. After explaining that it was not a hold-up, the observer measured the distance where the person had been sitting (back of chair to edge of desk). The mean distance was 19.1 inches, with a standard deviation of 4.8 inches. Only one of 32 people sat at 12 inches or less. It was observed, incidentally, that a strange sex difference seems to obtain in desk seating patterns. Nine of the 32 people sat at an angle to the desk (their distance was to the center of the chair back), and eight of these were male. The sample was equally divided between males and females. This difference did not seem to be task-related as all people except one were reading or writing—the one typist cannot explain the female tendency to sit with evenly placed chairs.

In the carefully designed inhospitality of the laboratory, the instructors measured how often and how much furniture was adjusted. The cubicles were examined after class, and the central area tables were examined after the pre-experimental discussion (while the students were involved in the experiment in the cubicles), after the mid-experiment switch in *E-S* roles (which necessitated students coming out of the cubicles and, often, through the frontal barriers, as well as switching chairs in the cubicles), and after class. The observation and measurement was done covertly. At these times, the inhospitable distances were reset when they had been adjusted and recorded.

Three instructors participated after explanation and training in the experimental procedure. Thirty-four students unwittingly served as *S*s. The following week they were informed of the experiment and queried as to their recollections of the experience.

Results

The two measures used, frequency and amount of furniture adjusted, were applied to two types of furniture, the central area tables and the cubicle chairs. Since the cubicle chair distance could only be increased by pushing the cubicle tables away, the data actually consist entirely of amount and frequency of table movement.

An estimate of total tight-squeeze passages was first made, in order to compare that number with central area table movements made. Each student had to enter and leave through the frontal barrier, and make the return trip once in mid-class for picking up role instructions in the memory experiment. In addition, each student had to make three entries and exits from a cubicle for the same reasons. Beyond that, students often emerged from the cubicles with a question, but as no count of the exact number of these queries was made, they are not included in the total. The total, a conservative one, is 238 passages through barriers no more than 6-7 inches wide (7 passages × 34 students).

The frontal barrier table gaps were adjusted by students exactly twice. The cubicle-entryway passages were adjusted twice by moving tables and three times by moving a cubicle wall panel. In all seven cases, adjustment just sufficient for passage without turning sideways was made. Thus, slightly more than 97% of all passages yielded to the position of the table and whatever else formed the other half of the tight gap. In less than 3% of all passages did students fail to accept this Scylla and Charybdis situation. And then they only moved the tables barely enough to squeeze through themselves. None of the 34 seemed remotely close to suggesting that the whole situation was uncomfortable or changing the room as a whole. Of course they had never received any direct communication that such behavior was not allowed.

The cubicle chair-table distances were, if anything, relatively tighter than the central area table distances, and a little more adjustment was observed. The mean adjustment from 12 inches was 1.9 inches. When the distribution, however, was skewed, 70% moved their distance 2 inches or less. Essentially, a few people moved the table quite a bit and most moved it not at all or only incidentally, perhaps accidentally. The most adjustment, to 17 inches, was done by three subjects. This is still 2 inches less than the mean of the naturalistic observation. The difference between the means of the 32 naturalistically observed people involved in deskwork and the 17 students' chairs (used by 19 students because some two of them switched chairs when the *E/S* roles were switched in the memory experiment) was significant ($t = 4.17, p < .001$).

Discussion

The data suggest quite strongly that students in a classroom will repeatedly (seven or more times) accept an im-

pediment rather than adjust it to levels of comfort. Most of these students also accepted, in the same space of about 90 minutes, an uncomfortable seating arrangement in cubicles. They spent a great majority of their sitting and walking time in the class experiencing and yielding to minor barriers of furniture. None of them made more than a short-range adjustment of tables and chairs to accommodate his or her own body at the moment, and even these subjects were very rare. One instructor-experimenter noted that two of the seven adjustments were a necessity—the student was simply too large to fit through a 7-inch gap.

All the observers noted student efforts to avoid moving the furniture, such as grunts, swiveling of hips, and willingness to line up for passage through a tight squeeze. The tables came to seem magically immobile; one knew they required only a tiny amount of effort to move, yet they withstood over 238 carefully maneuvered people-passages.

The following week when all students were told of the experiment and asked to recall their experience of it, surprisingly few (one) even remembered there being any form of impediment. Others were at a loss to recall it, although one volunteered the explanation that perhaps the tables were "supposed" to be that way. In their previous classes, tables and chairs were relatively disordered, with large handy gaps, as the author discovered when he began to set up a thorough system of impediments. If the present results are generally valid, one wonders how long it would take for students to get the tables disordered! (Of course maintenance workers might change table positions in the course of their duties.)

Why students adjust to furniture rather than adjusting it is not clear. The differences between the naturalistically observed deskwork situation and the experimental situation provide several hypotheses worth further investigation. Possibly, in student perception, institutionally owned furniture is not a part of the student's personal area of control. Yet the offices, where movement had been observed, also contained furniture not owned by the individuals. The differences which are salient are (*a*) that furniture is *perceived* as within personal control in an office and not in a classroom and (*b*) that the office is an individual (or perhaps a twosome) domain, while the class is a group of people. Probably the office group was an older group, and this indirectly or directly mediated the results. However, there is little doubt the experimental distances were below the comfort range for most people.

If task-involvement in the memory experiment, to the detriment of personal comfort, is advanced as a hypothesis, another implication arises. Though no check of student attitudes was made in this study, one would expect such repeated minor discomfort to develop into a variety of irritations and negative attitudes among the students. If they do not know why they feel badly toward a given class or situation, they are apt to ascribe it to whatever is most handy—the teacher, the school, their classmates. This could be the beginning of an unfortunate deterioration in whatever valuable relationships other efforts in the school had begun. The example used in this study, slight frequent altercations with tables, is not in itself significant; yet it may typify a range of subtle frustrations in classrooms which are below the threshold of awareness for all concerned. But if they are pointed out or discerned through a careful survey of the physical plant, even a new award-winning one (4), they can often very easily be changed or at least ameliorated.

FOOTNOTE

1. The author is grateful to Dr. Roger Blackman and Mr. Steve Richmond for their assistance as experimenters. This research was supported in part by grant W72-5414, The Canada Council. The experimental study was conducted at Simon Fraser University, Burnaby, B. C.

REFERENCES

1. "Comments on an Exhibition of Alex Colville, G. Smith, and Arthur Erickson," Simon Fraser University Art Gallery, May 1973.
2. Gifford, R.; and Sommer, R., "The Desk or the Bed?" *Personal and Guidance Journal*, 46:876-878, May 1968.
3. Proshansky, H., Introduction to H. Proshansky; W. Ittelson; and L. Rivlin (eds.), *Environmental Psychology*, Holt, Rinehart and Winston, New York, 1970.
4. Sommer, R., *Personal Space*, Prentice-Hall, Englewood Cliffs, N. J., 1969.
5. Sommer, R., *Design Awareness*, Rinehart Press, San Francisco, 1972.

AN A PRIORI APPROACH FOR DEVELOPING SHORT-FORMS OF TESTS AND INVENTORIES

JULIAN L. BIGGERS
Texas Tech University

ABSTRACT

The empirical item analysis techniques for developing short-forms of tests is questioned. Assumptions fundamental to Spearman-Brown theory are shown to offer an alternative. To test the assertions, a parallel short-form of Rokeach's Dogmatism Scale was developed. The resulting inventory is compared with two other short-forms produced using empirical methods of item selection. It was found that the *a priori*-developed scale generated satisfactory results. The method offers numerous advantages to the researcher.

A RESEARCHER WILL occasionally wish to measure a trait but then finds that the standard instrument for the measurement is too lengthy or time consuming to fit the design of a particular study. The obvious solution would be the development of an abbreviated or short-form of the parent instrument. All too often the researcher finds that the traditional process for developing a shortened scale involves more labor than his originally planned research, and he thus abandons measurement of the trait as part of the research. The purpose of this article is to point to a simplified procedure that may serve adequately for developing a short-form of a test when the basic assumptions are met.

The usual procedure for devising a short-form of a test is empirical in nature. The task involves (*a*) administration of the original instrument to a sample of examinees similar to the target population of the study; (*b*) item analysis to identify the items with highest association with the total score on the parent test; and (*c*) selection of the requisite number of items for the short-form. Additional work is required to establish the derived instrument's reliability and validity. Obviously, a great deal of labor and computation is required before the final product is obtained.

The desired end product of the effort is to produce a miniature parallel edition of the full-scale. If the parent instrument is viewed as being made up of *n* parallel short-forms, the task of the researcher becomes that of identifying one meeting the length requirements for his study. It should be obvious that the empirical approach does not accomplish this task. Items identified and selected with the highest item-total score correlations are likely to produce a short-form that is not parallel with other possible short-forms made up of the items not so selected. The empiricist develops a new short-form with some degree of concurrent validity with, but not parallel to, the full-scale. Applying the reverse process should make the point clear. Starting with the short-form and adding new items with similar characteristics to produce an instrument as long as the original full-scale should produce an instrument with statistical characteristics different from the original full-length edition. Cloaking an empirically produced short-form with all the characteristics of the parent instrument is a tenuous proposition at best.

The theoretical basis for the Spearman-Brown prophecy formula provides a simplified alternative to the empirical approach for developing a truly parallel short-form of a test. The Spearman-Brown formula is most often cited in the literature in association with estimating the reliability of lengthened tests. Overlooked is the practical aspect that the same assumptions apply when reducing the length of an instrument. Gulliksen (1) points out that when the assumptions are met, the Spearman-Brown formula provides exact results, not an estimate of the reliability obtained. Application of the Spearman-Brown formula requires that the test be unidimensional in the trait measured and that items removed be homogeneous with those retained. The procedure for producing a short-form via Spearman-Brown theory would only involve (*a*) obtaining or assuming the reliability of the full-scale for the population to be examined; (*b*) estimating the reduction in scale length to produce an acceptable lower reliability for the research instrument; and (*c*) selection of the items to retain (or remove) to produce the short-form.

The *a priori* method has obvious advantages over the empirical method. Only an estimate of the reliability of the total scale for the target population is needed. This requirement may occasion the administration of the instrument to a sample, but that task is inherent in both procedures. When the reliability is already known for similar samples, the researcher may be willing to accept this estimate without further effort. All remaining work for estimating the reduced length or reliability and the

sampling procedure to select items can be carried out in a few moments at the researcher's desk. Finally, the resulting short-form will have reliability and length characteristics dictated by the researcher in advance. Using the empirical approach, only the length might be determined in advance; the reliability would have to be obtained after actual use of the short-form.

Adherents to the empirical method may concede that the foregoing is true, but will likely point out that although their way is arduous, expensive, and time consuming, the method results in the identification and selection of items most appropriate for the population to be tested. The assertion is not denied, but it may, as discussed previously, be overstressed. A high coefficient of internal consistency suggests a strong item-total score correlation, and the Spearman-Brown approach obtains a sample of those items to produce a parallel form with a predetermined reliability. Neither approach, empirical or *a priori*, would be appropriate if the full-scale had a low reliability coefficient for the population.

The Problem

The basic question to be answered is, How will a short-form produced by the *a priori* Spearman-Brown method compare with an empirically developed instrument? To obtain an answer, a short-form of Rokeach's Dogmatism Scale (2) was produced and compared with two empirically produced versions.

Schulze (3) developed a 10-item version of the Dogmatism Scale using Guttman's scalogram analysis as the mode for empirically selecting the items. Troldahl and Powell (4) produced data for a 20-item scale and also for even shorter versions using the regular item analysis technique. Schulze tested college students in his work, while Troldahl and Powell studied adult residents of Lansing, Michigan, and Boston. The two empirically developed scales have only four items in common demonstrating the variable item values obtained in different populations and with different item-selecting techniques. Neither study reports direct evidence of reliability of the short-form in the usually accepted sense. Schulze reported a coefficient of reproducibility (.83) and pointed to similar results obtained in two separate studies as evidence of the reliability of his 10-item scale. Troldahl and Powell used a statistical procedure involving variance estimates (described later) to conclude that their 20-item instrument would have a split-half reliability of .79.

Methodology

A short-form of the Dogmatism Scale was developed by selecting the twenty odd-numbered items from Rokeach's 40-item, Form E, scale. The item-selection method will be recognized as the traditional split-half technique. It meets requirements of the assumptions and is a simpler selection process than the more complicated randomized item-selection procedure. The resulting *a priori*-developed scale, using 50% of the available items, had close to a chance level commonality with the two empirically produced scales. Four of Schulze's ten items (40%) and nine of the twenty items (45%) in the Troldahl and Powell scale appeared among the odd-numbered items selected for the experimental short-form.

The twenty odd-numbered items from the Dogmatism Scale were merged with twenty items from the F-Scale to produce an instrument similar in length to the full Dogmatism Scale. The experimental short-form and the full-length Dogmatism Scale were administered to undergraduates enrolled in an educational psychology course. A two-week interval occurred between administrations. Scores were obtained for the experimental short-form, the full-length inventory, and the odd- and even-numbered halves of the full-scale. Product-moment correlations were computed between the four sets of scores thus obtained. Table 1 summarizes the results of the analysis.

Table 1.—Intercorrelations of the Short-Form and Full-Length Dogmatism Scale

Test Forms	Correlations between Test Forms			
	A	B	C	D
A. Short-Form (Experimental)	1.00	.75	.78	.61
B. Full-Length Scale	.75	1.00	.92	.93
C. —Odd-numbered Items	.78	.92	1.00	.71
D. —Even-numbered Items	.61	.93	.71	1.00

Analysis of Results

The correlation between the odd and even halves of the full-length Dogmatism Scale when inserted in the Spearman-Brown formula produced an estimated reliability coefficient of .83 for the full scale. This value is within the range of reliabilities reported by Rokeach (2: 89-90). The correlation of .78 between the experimental short-form and the odd-numbered items of the full-scale may be interpreted as the test-retest reliability for the abbreviated scale. In a similar fashion, the correlation of the experimental short-form with the even-numbered items could be considered an estimate of the alternate form reliability after a two-week interval. Lastly, the experimental form's correlation of .75 with the full-length version is the estimated predictive validity coefficient. The reliability estimates (split-half, test-retest, and alternate form) all appear to be in an acceptable range to warrant use of the short-form in group studies. The predictive validity coefficient is as high as that found in many studies of this nature.

The statistical procedure used by Troldahl and Powell was followed in obtaining reliability estimates for all three instruments for comparison purposes. The full-length Dogmatism Scale in this study had a corrected split-half reliability of .83, which indicates that approximately 69% of the total variability is explained by the attributes the items had in common. A correlation of .92 was obtained between the odd-numbered items and the

full-scale. This indicates the 85% of the variability in the full-scale is explained by the odd-numbered items. An estimate of "true" variability represented by the odd-items is 58% (.69 × .85). The square root of this percentage is an estimate of the split-half reliability of the experimental short-form which is .77. When Troldahl and Powell followed this procedure, they obtained an estimated split-half reliability of .79 for their empirically derived 20-item scale. The 10-item Schulze scale has an estimated reliability of .62 applying the data available in a similar fashion. Attenuated to double length for comparison sake, an estimated reliability of .76 was obtained.

All three short-forms have comparable reliability estimates. The slight variations might be attributed to rounding error or differences in reliability of the original full-scale for the populations tested.

Summary

The use of the Spearman-Brown theory to produce a short parallel version of a scale seems to be warranted. The parent instrument should fulfill the requirements of assumed unidimensionality and general equivalence of intercorrelations of items to be eligible. In addition, the reliability of the full-scale should be high enough for the popularion to be surveyed to allow for reduction by shortening.

A short-form of Rokeach's Dogmatism Scale was produced by selecting the twenty odd-numbered items. The *a priori*-developed scale was shown to have satisfactory statistical properties for a half-length edition. The experimental scale had a reliability coefficient similar to that of two empirically developed scales even though item overlap between the scales was near the chance level.

REFERENCES

1. Gulliksen, Harold, *Theory of Mental Tests,* John Wiley & Sons, New York, 1950.
2. Rokeach, Milton, *The Open and Closed Mind,* Basic Books, New York, 1960.
3. Schulze, Rolf H. K., "A Shortened Version of the Rokeach Dogmatism Scale," *Journal of Psychological Studies,* 13:93-97, 1962.
4. Troldahl, Verling C.; and Powell, Fredric A., "A Short-Form Dogmatism Scale for Use in Field Studies," *Social Forces,* 44:211-214, 1965.

PEER JUDGMENTS OF TEACHING COMPETENCE AS A FUNCTION OF FIELD INDEPENDENCE AND DOGMATISM[1,2]

JAMES B. VICTOR
State University of New York at Albany

ABSTRACT

The theoretical relationships among field independence, dogmatism, and a peer judgment criterion of professional competence were examined. The subjects were master's level interns in a training program. The data illustrate the reliability of the criterion and indicate that these interns differentiate between professional competence judgments and more interpersonal judgments when making peer choices. Neither dogmatism nor field independence alone predicts the criterion, but the interaction term for the two variables significantly predicts the criterion. It is the field dependent/highly dogmatic person who is chosen less often by his peers, while the field dependent/low dogmatic person is chosen more often.

IN RECENT YEARS considerable attention has been paid to the issue of competence criteria for teachers, social workers, and others who work with children. Several authors have indicated peer judgments of preferred work-partners, in training situations, to be related to certain interpersonal qualities, such as self-disclosure and interpersonal flexibility (10) and adaptable teaching behavior (11). The present study was designed to expand the network of such interpersonal variables.

The program for training teachers of emotionally disturbed children at SUNY Albany is well-suited for gathering sociogram data. For one semester, each intern

works closely with six to eight other interns in one of three settings for emotionally disturbed, neurologically impaired, or behaviorally difficult children. The interns share responsibility for program planning, case conferences, and the day-to-day workload. In addition to the intense daily work contact, the interns are together during academic course work. This setting provided the opportunity to discern who would come to be judged the more competent interns in the group.

Witkin (21) and his colleagues have described a dimension which they call *field independence.* This variable differentiates individuals in terms of their active striving, analytic attitude, and degree of self-awareness. In terms of interpersonal functioning, Witkin (20) states that "the less developed sense of separate identity of persons with a global cognitive style manifests itself in reliance on external sources for definition of their attitudes, judgments, sentiments, and of their view of themselves." A number of studies have illustrated that field dependent persons are more prone to be guided by positions attributed to an authority figure or peer group (1, 3, 13), remember verbal messages that are more social in context (4, 5, 6, 7, 8), and adapt their performance on a cognitive task to a modelling demonstration viewed on TV (19). While these data do clarify a person's characteristic interpersonal style, they do not give information as to the way the person will be viewed by others.

Rokeach (18) describes another dimension of cognitive style which has been viewed as important in interpersonal functioning. This construct, *dogmatism,* is thought to be related to a person's openness to new ideas and to the independent evaluation the person is able to make on incoming information. Dogmatism and field independence have both been seen as important constructs in teachers' interpersonal functioning. Measures of dogmatism and field independence share little variance with each other and display very low correlations with measures which purportedly assess open, other-centered attitudes and behavior. Clearly the views of those who define interpersonal functioning in terms of relative isomorphism between interpersonal openness and such constructs as dogmatism and/or field independence are overly simplistic.

Since field independence and dogmatism are essentially uncorrelated, it is possible to identify individuals representing combinations of levels on both constructs. Several studies employing this conceptualization have found that it is the high dogmatic/field dependent person who tends to be different from others (14, 15, 16, 17). In these studies the high dogmatic/field dependent person has been found to have difficulty with both reversal and non-reversal shift concept-formation problems, score low on inventory scales of predictive surgency or dynamism of classroom teaching behavior, and score lower on a creativity test.

It was the aim of this study to put the Ohnmacht (14) conceptualization to a direct test of whether field independence and dogmatism taken together were related to peer judgments in a teaching situation. The hypothesis that the interaction term of field independence and dogmatism would be related to peer judgments of competence was drawn from the earlier Ohnmacht studies. In particular, persons with combinations of field dependent /high dogmatic scores were viewed as most likely to receive a *lower* number of nominations. Also, because of their reliance on others for self-definition, field dependent interns who scored low on dogmatism were predicted to receive a *higher* number of nominations.

Method

Subjects

The *S*s were 50 master's level students in an intern training program for teachers of emotionally disturbed children. All students accepted into the program for a two-year period were included in the study. Program selection was made by usual procedures of test scores, previous academic records, and interviews. Selection staff were unaware of the students' scores for both of the variables used in this study. All *S*s were enrolled, after selection, in the same internship teaching practicum.

Procedure

Before the program began, all *S*s were administered the Hidden Figures Test (HFT), a measure of field independence (12), and the Dogmatism Scale (DS), a measure of openmindedness or dogmatism (18).

After one semester of practicum experience each S was asked to nominate other individuals in his work group as his first, second, or third choice as a partner for various activities, that is, the person with whom *S* would prefer to (*1*) teach emotionally disturbed children; (*2*) develop programs programs for emotionally disturbed children; (*3*) work with as a consultant in regard to emotionally disturbed children; *(4)* talk to about a personal problem; and *(5)* take to a party.

The inter-judge reliability was determined for these five questions for each group using a formula provided by Gordon (9). The judgments for "teach," "consult," and "develop programs" were very similar, with reliability coefficients each ranging from .49 to .95 with the median at .74. The judgments for "take to a party" and "talk to about a personal problem" were less reliable, ranging from .32 to .50 with the median at .42.

The factor analysis using varimax rotation yielded two orthogonal factors, The first, labelled *professional competence*, showed loadings $> .83$ for the peer choices of teaching, consulting, and program developing. The second factor, *interpersonal-social*, loaded $> .85$ for choices of taking to a party and talking to about a personal problem. The correlation matrix and factor loadings are illustrated in Table 1. Two new individual difference variables were formed using factor scores for *professional competence* judgments and *interpersonal-social* judgments.

Table 1.—Correlation Matrix and Factor Loadings for Sociogram Items

Sociogram Items	Correlation Matrix					Factor*	
	1	2	3	4	5	I	II
Develop program	—	.73	.62	.42	.30	.87	.24
Teach		—	.55	.30	.39	.84	.23
Consult			—	.20	.34	.83	.11
Personal problem				—	.59	.16	.89
Take to party					—	.22	.85

*Factor I accounts for 56% and Factor II 21% of the total variance

Table 2.—Correlation Matrix for Variables

Variable	Correlation Matrix				
	1	2	3	4	5
Professional competence	—	−.05	−.13	.18	.41*
Interpersonal-social		—	.03	−.12	−.07
Dogmatism			—	−.15	.19
Field independence				—	−.08
HFT X DS					—

*$p < .01$

A regression analysis (2) was performed on five final variables. The independent variables were HFT, DS, and HFT X DS, the interaction term. A separate analysis was performed against each of the criteria, *professional competence* and *interpersonal-social* judgments.

Results

The correlation matrix for the five final variables used in the study yielded one correlation with a statistically significant value (*professional competence* and HFT X DS: $r = .41, p < .01$) as can be seen in Table 2.

The full model regression analysis yielded a significant effect for the criterion *professional competence* judgment, as can be seen from Table 3 ($F = 5.09$, $df = 3/46$, $p < .005$). The main effects of HFT or DS did not reach levels of significance; however, the HFT X DS interaction term was statistically significant ($F = 12.53$, $df = 1/46$, $p < .005$). A regression analysis of the *interpersonal-social* judgment criterion did not yield even nominal levels of significance.

A median split technique was applied to HFT and DS scores to test the hypothesis that persons scoring low on HFT and high on DS would receive fewer peer choices than those scoring low on both HFT and DS. The score used for each S was his average nomination for the choices of teaching, consulting, and developing programs. The means and standard deviations for each of the four cells are presented in Table 4. The difference between the low HFT-high DS and the low HFT-low DS groups yielded a statistically significant quantity ($t = 1.96$, $df = 24$, $p < .05$, one-tailed test).

Discussion

The teachers in this study did differentiate between *professional competence* judgments of their peers, which were more reliable, and *interpersonal-social* judgments. The work conditions of the groups did vary in the amount of close contact that each intern had with his fellow workers, and this seemed to affect the inter-judge reliability estimates. In general, the closer the contact during the work experience, the higher the correlation coefficient.

The present data are consistent with the Ohnmacht studies (14, 15, 16, 17) that persons who score high on the dogmatism scale and are field dependent are the most predictible group, and they provide weak support for the idea that these variables when considered together provide useful information about which teachers will be valued by their colleagues as professionally competent. Interns with the particular combination of high dogmatism/field dependent scores were chosen less by their peers, while those who had low dogmatism/field dependent scores were chosen more often. It is noteworthy that the only three interns in the sample who were viewed as isolates by their peers were in the high dogmatism/field

Table 3.—Multiple Regression Results for the Professional Competence Peer Judgment Criterion

Predictor	R^2	F	df	Partial R
Dogmatism		2.13 n.s.	1/46	−.21
HFT		2.22 n.s.	1/46	.21
HFT X Dogmatism		12.53*	1/46	.46
Full model	.25	5.09*	3/46	

*$p < .005$

Table 4.—Average Peer Selection Scores for Choices of Teaching, Consulting, and Developing Programs for *Ss* in Cells, Using Median Split Technique on HFT and DS

Field Independence	*Dogmatism*					
	High			*Low*		
	Mean	*SD*	*N*	Mean	*SD*	*N*
High	2.67	2.10	11	2.34	1.23	14
Low	1.83	1.56	14	3.00	1.29	11

dependent group. In summary, the study confirms the hypothesis that the personal characteristics of dogmatism and field dependence contribute to competence judgments. However, it is the interaction of these variables that is the determining factor.

FOOTNOTES

1. Portions of this paper were presented at the Meeting of the Eastern Psychological Association, Washington, D. C., 1973. The author wishes to thank Dr. Oliver M. Nikoloff for providing valuable assistance in support of this study.

2. Requests for reprints should be sent to the author's address: Department of Educational Psychology and Statistics, State University of New York at Albany, 1400 Washington Avenue, Albany, N. Y., 12222.

REFERENCES

1. Bell, D. R., "The Relationship between Reward and Punishment-Avoidance Orientation and Selected Perceptual Variables," unpublished doctoral dissertation, University of Oregon, 1964.
2. Cohen, J., "Multiple Regression as a General Data-Analytic System," *Psychological Bulletin,* 70: 426 - 443, 1968.
3. Deever, S. G., "Ratings of Task-Oriented Expectancy for Success as a Function of Internal Control and Field Independence," unpublished doctoral dissertation, University of Florida, 1967.
4. Eagle, M.; Fitzgibbons, D.; and Goldberger, L., "Field Dependence and Memory for Relevant and Irrelevant Incidental Stimuli," *Perceptual and Motor Skills,* 23: 1035 - 1038, 1966.
5. Eagle, M.; Goldberger, L.; Breitman, M., "Field Dependence and Memory for Social vs. Neutral and Relevant vs. Irrelevant Incidental Stimuli," *Perceptual and Motor Skills,* 29: 903 - 910, 1969.
6. Fitzgibbons, D.; and Goldberger, L., "Task and Social Orientation: A Study of Field Dependence, 'Arousal,' and Memory for Incidental Material," *Perceptual and Motor Skills,* 32: 167 - 174, 1971.
7. Fitzgibbons, D.; Goldberger, L.; and Eagle, M., "Field Dependence and Memory for Incidental Material," *Perceptual and Motor Skills,* 21: 743 - 749, 1965.
8. Goldberger, L.; and Bendich, S., "Field Dependence and Social Responsiveness as Determinants of Spontaneously Produced Words," *Perceptual and Motor Skills,* 34: 883 - 886, 1972.
9. Gordon, L. V., "Estimating the Reliability of Peer Ratings," *Educational and Psychological Measurement,* 29: 305 - 313, 1969.
10. Halverson, C. F.; and Shore, R. E., "Self-disclosure and Interpersonal Functioning," *Journal of Consulting and Clinical Psychology,* 33: 213 - 217, 1969.
11. Hunt, D. E., "Adaptability in Interpersonal Communication among Training Agents," *The Merrill-Palmer Quarterly of Behavior and Development,* No. 16, 1970.
12. Jackson, D. N.; Messick, A.; and Myers, C. T., "Evaluation of Group and Individual Forms of Embedded Figure Measures of Field Independence," *Educational and Psychological Measurement,* 24: 177 - 192, 1964.
13. Linton, H.; and Graham, F., "Personality Correlates of Persuasibility," in I. Janis *et al.* (eds.), *Personality and Persuasibility,* Yale University Press, New Haven, 1959.
14. Ohnmacht, F. W., "Effects of Field Independence and Dogmatism on Reversal and Nonreversal Shifts in Concept Formation," *Perceptual and Motor Skills,* 22: 491 - 497, 1966.
15. Ohnmacht, F. W., "Teacher Characteristics and Their Relationships to Some Cognitive Styles," *Journal of Educational Research,* 60: 201 - 204, 1967.
16. Ohnmacht, F. W., "Factorial Invariance of the Teacher Characteristics Schedule and Measures of Two Cognitive Styles," *Journal of Psychology,* 69: 193 - 199, 1968.
17. Ohnmacht, F. W.; and McMorris, R. F., "Creativity as a Function of Field Independence and Dogmatism," *Journal of Psychology,* 79: 165 - 168, 1971.
18. Rokeach, M., *The Open and Closed Mind.* Basic Books, New York, 1960.
19. Toomey, T. C.; "Alteration of a Perceptual Mode Correlate through a Televised Model," *Journal of Experimental Research in Personality,* 6: 52 - 59, 1972.
20. Witkin, H. A., "Psychological Differentiation and Forms of Pathology," *Journal of Abnormal Psychology,* 70: 317 - 336, 1965.
21. Witkin, H. A.; Dyke, R. B.; Faterson, H. F.; Goodenough, D. R.; and Karp, S. A., *Psychological Differentiation,* John Wiley, New York, 1962.

CAN SUGGESTIONS BY TEACHERS IMPROVE INSTRUCTION?

JOHN D. McNEIL
University of California, Los Angeles

ABSTRACT

Twenty-four experienced teachers each selected a different reading skill and each prepared a set of written suggestions for teaching the skill selected. Subsequently, 24 unsuccessful teachers, i.e., those less able than their peers to effect pupil achievement in the skills, were identified. These unsuccessful teachers were randomly assigned to two treatment groups. One group received the previously prepared suggestions and the other group did not. Later, the two groups of teachers again taught the skills. It was found that all teachers receiving suggestions improved relative to their previous performance, i.e., their pupils achieved more, while only slightly more than half of those without suggestions showed improvement ($p < .05$). Thus, it was concluded that suggestions for teaching can be helpful to less successful teachers.

TEACHING HAS BEEN a very private kind of work. Customarily, teachers have worked in different rooms at the same time, thereby making it difficult to see each other teach and, accordingly, impossible to help each other on the basis of direct observation. Furthermore, as indicated by Dreeben (2), teachers have lacked written media for communicating about their work because their occupation has had no counterpart to the scholar's research tradition in which knowledge is accumulated in books and journals, or to the physician's case records in which tests and prior medical decisions are documented.

Now, however, there are signs that the fragmentation of the colleague group may diminish and that teachers may not be left alone to determine what they are doing right or wrong. For example, new guidelines of the Right to Read program stress staff development involving all school personnel in activities directly related to everyday classroom instruction (1). Team teaching, diversified staffing, videotaping of lessons, faculty intervisitations, and teachers' centers are other innovations that promise to make it easier for teachers to aid one another in school settings as they work through their instructional problems.

Although it is a well-substantiated fact that teachers continue to receive most of their assistance for self-recognized weaknesses from their peers or from their own trials and success (3), there is little evidence that the help received makes a difference on the growth of pupils. We do not know whether suggestions from peers are valid, i. e., contribute to pupil achievement. Indeed, teacher preference for peer assistance might be nothing more than a defense against such options as (*a*) supervisors who threaten to reveal inadequacy; (*b*) unrealistic college experts freed from the demands of real-life children; and (*c*) administrators who are sure to remember a revealed weakness when completing an annual evaluation form.

The purpose of this study was to determine whether teachers could provide suggestions that would improve the ability of other teacher to effect pupil progress. The study was constituted to provide for wide generalization within the field of reading; i. e., a large number of suggestions were made for teaching many skills of reading at different grade levels and these suggestions were given to teachers with varied backgrounds. The following design factors were employed to maximize the possibility ot finding value in teachers' suggestions: (*a*) Suggestions were specific to the teaching of particular reading skills which were operationally defined; (*b*) Each teacher who provided the suggestions for a given teaching task was very familiar with that task; and (c) The population of teachers within which the effect of suggestions was to be noted consisted of teachers who needed help because they were inferior to their peers in teaching the particular skills.

Method

Subjects

The "suggestors" were 24 teachers who provided written suggestions. They were all experienced teachers in the Los Angeles area who were candidates for a Master's degreé in the teaching of reading. Their ages ranged from 22 years to 56 years, and they were characterized as highly verbal, scoring above 48 on the Miller Analogies Test.

The teachers selected to receive or not receive suggestions, i. e., the experimental or control teachers, were those whose pupils did not achieve under their direction as well as other pupils taught by other teachers. These unsuccessful teachers had a wide range in background. Some of them had taught for more than 14 years, others

were beginning teachers, and a few were instructional aides.

Tasks and Materials

Each of the 24 suggestors also provided a teaching performance task. Each task was in the form of a mini-lesson consisting of a measurable instructional objective, a sample test item, information regarding the importance of the objective as a reading skill, and background information that a teacher could use in planning the lesson. It was made clear that the person who would teach the task was free to design her own lesson within the constraints of the objective. Each task was intended for a particular population of learners, e. g., children in kindergarten, high school students in remedial reading. It was specified that the teacher should have 30 minutes for preparation and 15 minutes for teaching the lesson. The objectives of the tasks included recognizing open and closed mouth sounds; identifying words that rhyme; matching initial sounds of spoken words with pictures whose names begin with the same sound; discriminating vowel sounds of different printed words; applying the "final *e*" rule; arranging pictures in sequential order according to a story; identifying compound words; recognizing spelling patterns and using them in decoding new words; identifying statements of fact and opinion; differentiating specific and general words; using guide words; distinguishing homonyms; interpreting metaphors; identifying thesis sentences.

Prior to preparing their suggestions for teaching the different lessons, the suggestors made task analyses for themselves in order to identify the prerequisites, that is, to see what was involved in the task. They also constructed 10-item criterion referenced tests with which to assess pupil attainment of each objective. In most instances, these 24 teachers composed and tried out their own lessons to ensure that the tasks were appropriate for intended learners and to gain confidence in the procedures which they would suggest to others.

Written Suggestions

The suggestions were task specific. General principles, when given, were followed by examples of how the general term applied in the particular case. Suggestions were written for each of the following principles: perceived purpose; motivational appeal, e.g., use of color, learners' personal experiences, humor; use of multiple active responses, including manipulation; teaching prerequisite skills; sequencing from simple to complex and from known to unknown; use of mnemonic devices; inductive and deductive methods; appropriate practice; analogous practice; knowledge of results; prompting, praise; review; avoiding both irrelevant practice and attending to one child to the loss of others.

By way of example, the suggestions that were given for teaching the mini-lesson "recognizing compound words" appear below:

1. Write several compound words on the board. Let the children look at the words for a moment. Welcome inquiries or deductions regarding the words.
2. Ask children what is special, common, etc., to all the words.
3. If necessary, prompt the children to recognize that the words on the board have words within words—two words put together. Reinforce any comments made with regard to two words in one, words put together, etc.
4. Ask children if they know what these words are called.
5. Define a compound word. Ask children if they can think of some compound words on their own. Ask children to identify the words that made up the compound words supplied.
6. Try to distract the children by mixing compound words with words that contain prefixes, suffixes, and poetic prepositions. Contrasting prefixes and suffixes to components of compound words is always insightful.
7. Write a list of words on the board and have children come up and circle the compounds and divide their parts with a slash; or pass out a little quiz and have children complete it at their desks. Correct the exercise together immediately.
8. Problems: If children mix up compounds with base words with prefixes and suffixes, stress the fact that compound words are made up of more than one word. Contrast this with the prefix or suffix which is a tag-along. For example, can the two parts of this word stand by themselves? *preview—pre/view.* Can the two parts of this word stand by themselves? *moreover—more/over.*

Procedure

The suggestors went with their performance tasks and suggestions to schools where they each selected two teachers with pupils at a level appropriate for the task. Each pair of teachers selected was asked to undertake a performance task, teaching a particular reading skill to a group of six or eight pupils. The pupils were to be randomly chosen from those present and named on the class register. Each member of each pair of teachers was given 30 minutes to design her own individual lesson, adhering to the same instructional objective as her peer. The suggestor observed the 15-minute lessons which ensued and then administered post-tests to the pupils in order to assess the effects of the instruction.

Later a coin was tossed to decide which teacher in each pair would receive suggestions before teaching another group of children the same reading skill. This experimental teacher was asked to redesign her original lesson to include the suggestions; the control teacher was asked to teach the lesson a second time to different pupils "in whatever way you think best." Both teachers in each of the 24 pairs were again allowed 30 minutes for their preparation. Children for the second lessons were randomly drawn from those present in the rooms who had not been taught before. Lesson observations and post-testing were conducted as in the first trial.

Analysis

It is recalled that each group of two teachers taught to a different objective. The high-achieving teacher in each

pair was the one whose pupils earned a higher group mean score. This teacher was designated "successful." The teacher whose pupils did not score as well was designated "unsuccessful." The purpose of the study was to assess the effect of suggestions only upon teacher who needed help—the unsuccessful ones. Therefore, the unsuccessful teacher in each pair was categorized as either experimental (one who received suggestions) or control (no suggestions given). There were 12 experimental teachers and 12 controls, all of whom were originally declared unsuccessful because of poorer performance relative to a peer.

The question was whether or not these unsuccessful teachers improved on a second trial with or without suggestions. A teacher was said to have improved if the group mean score of her pupils following the second lesson was higher than the group mean score obtained from the first lesson.

A chi-square test with Yates' correction was applied inasmuch as the categories of improved or not improved teachers were based on different scores taken from independent samples of subjects.

Results

The overall results of the significance test for the effect of teacher suggestions upon the improved performance of unsuccessful teachers are summarized in Table 1.

Table 1.—Summary of the Significance Test for the Effect of Teacher Suggestions upon the Performance of Unsuccessful Teachers

Group	Improvement	No Improvement
With suggestions	12	0
Without suggestions	7	5

$\chi^2 = 4.04 \quad p < .05$

All teachers with suggestions improved relative to their first performance; slightly more than half of the teachers without suggestions did better. It is interesting to note that 58% of the teachers with suggestions exceeded their high-performing peers on the second lesson, while only 25% of the teachers without suggestions were able to do so.

Discussion

The results imply that the suggestions of teachers are helpful to less successful teachers. It is difficult to show the effect of suggestions upon high-achieving teachers. This is so because the high achievers already are likely to be using many of the suggested principles. Also, the possibility of improvement is more remote when one is already excelling.

The conditions by which teacher suggestions were formulated and given in this study may be regarded as atypical. Teachers seldom have the time to analyze carefully what is involved in the teaching of a particular objective and to design ways to maximize its attainment. Further, few teachers set teaching tasks for colleagues and give specific directions for teaching these tasks. However, the study should not be discounted for such irregularity.

Staff development might be enhanced by encouraging teachers to study systematically the instructional tasks that they and their peers believe to be important. The idea of teachers attempting to validate the suggestions of peers could give rise to more alternative modes of teaching and, thereby, serve more pupils effectively. If improvement can come when teachers follow suggestions for teaching involving an imposed task, a task which one does not necessarily regard as crucial, imagine what the results might be when suggestions are directed at those objectives considered vital by the teachers.

There are at least two factors that keep suggestions by peers from helping the less successful teacher. One, the suggestions themselves may be of little value, so weak that they cannot effect pupil gain. Two, the teacher might not be willing or able to apply the suggestions even if the suggestions are powerful. The fact that these factors seemed not to dominate in the present study offers hope that teachers can one day share their cumulative knowledge, thereby making teaching both more effective and a less solitary kind of work.

REFERENCES

1. Department of Health, Education and Welfare, "Right to Read Program," *Federal Register, 39,* 161: 29929-29931, August 1974.
2. Dreeben, Robert, "The School as a Workplace," *Second Handbook of Research on Teaching,* Rand McNally & Co., Chicago, 1973, pp. 450-471.
3. Lewis, Arthur J.; and Miel, Alice, *Supervision for Improved Instruction,* Wadsworth Publishing Co., Belmont, Calif., 1972, p. 190.

USING AN ACADEMIC PEER INTERACTION CONTINGENCY WITH EMOTIONALLY DISTURBED CHILDREN[1]

ROBERT C. COON
VINCENT A. ESCANDELL
Louisiana State University, Baton Rouge

KATHRYN B. COON
JULIET C. GREEN
East Baton Rouge
Parish School Board

ABSTRACT

An academic peer interaction contingency was introduced into an ongoing token economy program in a class of five emotionally disturbed children with minimal interpersonal skills. Seven behavioral categories of academically relevant and irrelevant behaviors were recorded during multiple baseline, contingency, and follow-up observation periods. An ANOVA revealed a significant category effect and a significant category by time interaction, indicating significant changes in the distribution of student behavior across categories as a function of contingency introduction. Hypothesized increases in student academic cooperation with peers and teacher, and hypothesized decreases in student academic work alone occurred at statistically significant levels. It was concluded that enduring academically constructive changes in interpersonal interaction within the classroom occurred as the result of contingency introduction.

TOKEN REINFORCEMENT PROGRAMS have been used in recent years to increase academically appropriate behaviors in a wide variety of subject populations, including emotionally disturbed children (6). The majority of such programs have had as a primary goal the decrease of disruptive classroom behavior (5, 9). In these programs children typically earn tokens by engaging in academically appropriate behaviors such as remaining in their seats, raising their hands, and doing work correctly (2, 5).

Despite the fact that emotionally disturbed children experience considerable behavioral difficulty related to academic peer interaction in the classroom setting, few token economy programs with this subject population have emphasized peer academic interactive categories of appropriate behavior. The few studies that have focused on the peer interaction of emotionally disturbed children within the classroom (1, 7) have generally been concerned with individual children and their relation to a group of non-problem children in the regular classroom. In addition, such studies have usually employed indirect methods of reinforcing peer interaction by increasing the attractiveness of the target child (4).

The present study extended previous token economy research by modifying an ongoing token economy program with emotionally disturbed children. In addition to the usual goal of increasing individual academically relevant behavior, the modified program reported in the present study sought to increase systematically positive peer academic interaction within the classroom. Since academically relevant behaviors constituted the target behaviors in the present study, it was hypothesized that students would show a significant increase in academic cooperation with both students and teacher. In addition, academic work alone was expected to decrease significantly.

Although additional behavioral categories were observed (e. g., "negative behavior," "other behavior") in order to provide a complete picture of each *S*'s behavior during the course of the study, these dealt with non-academic behaviors. While change in these categories was obviously anticipated since the category system was exhaustive in terms of each *S*'s behavioral repertoire, it was not possible to anticipate the specific ways in which these categories would be affected by the manipulation. Consequently, no specific directional hypotheses were made for the non-academic categories.

Method

Subjects

The *S*s were five black children of lower socioeconomic status in a special education class for the emotionally disturbed. Recent WISC Full Scale IQ scores ranged from 70 to 90. The group consisted of four males, ages 6, 10, 10, and 11, and one 8-year-old female. Data from an additional male class member, age 6, is not reported since he attended class infrequently.

On the basis of modal behavioral characteristics, the students were divided into two groups: Group 1 contained one female and the youngest male, who both were withdrawn and rejected by all class members; Group 2, consisting of the older students, contained three males who were outgoing and socially accepted by all class members. All students had been class members for at least five months prior to the present study.

Procedure

The *Ss* became involved in a token economy system upon admittance to the class. Consequently, all *Ss* had at least five months' experience with an ongoing token economy system in which points were earned for a variety of behaviors prior to the program modification described below. In the ongoing token economy program *Ss* could earn 1 point by maintaining appropriate behavior for each of 7 thirty-minute periods during the school day. Additional points could be earned during any thirty-minute period by completing an assignment, playing a game, listening to a record, watching a filmstrip, completing an art project, or completing other appropriate classroom activities. The *Ss* usually earned a minimum of 30 points for an average day.

Ss had the option of accumulating their points over time or spending them at the end of the school day. Back-up tokens such as toys, candy, games, and other objects were available to be purchased at the *Ss*' choice, at exchange rates which varied from 3 points (permitting daily redemption) to 500 points (requiring accumulation of points over several days or weeks).

Since the manipulation described in the present study was initiated as a treatment innovation in an ongoing applied classroom setting, a within-subject multiple AB design with follow-up observations was employed (6). Once the effective behavioral changes described below appeared as a result of introducing a peer academic interaction contingency into the ongoing token economy program, it was not feasible to return to precontingency conditions in this particular classroom setting.

For the present study a peer academic interaction contingency was introduced following a baseline observation period. All *Ss* were told that if they worked with a person who was not in their group they would obtain an extra point for each activity completed with that person. Thus, if a unit of work was completed with a child in the other group, 2 points were earned instead of the usual 1 point. To earn the extra point a *S* had to be engaged in an active working partnership with a member of the other group, rather than merely sitting with each other.

The *Ss*' behaviors were recorded during baseline, contingency, and follow-up observation periods by an observer in the classroom. The observer was familiar to the children and had frequently been present in the classroom as a nonparticipant in class activities prior to as well as during the study. All observation periods were fifteen minutes long. The dependent measure consisted of the mean percentage of time that each *S* was engaged in each of seven behavioral categories during multiple baseline, contingency and follow-up observation periods. The behavioral categories, which were designed to be mutually exclusive, were as follows:

1. Academic work alone
2. Cooperative academic interaction with another student
3. Cooperative academic interaction with teacher
4. Cooperative non-academic interaction with another student
5. Cooperative non-academic interaction with teacher
6. Negative interaction (fighting, verbal abuse of another, etc.)
7. Other behaviors (e. g., sitting alone while not engaged in academic work or interacting with others; aimlessly wandering around the room)

The first three categories were defined as having academic material on the desk for thirty or more seconds during each minute of each fifteen-minute observation period. The student was required to be actively involved with the materials (e. g., writing, turning pages) for the categories to be scored.

The next three categories were defined as any type of verbal or physical exchange with others while not engaged in an academic activity.

Pre-testing indicated that these categories could be used reliably by more than one observer and could account for all of a *S*'s behavior during observation periods.

Data were recorded over a three-week period. A total of 55 fifteen-minute time blocks was used during the three-week period. Baseline data, consisting of 17 random fifteen-minute observations was gathered during the first three days prior to contingency introduction. Contingency data, consisting of 38 random fifteen-minute observations, were gathered during the remaining twelve days. Follow-up consisted of 5 random fifteen-minute observations obtained more than one month later and distributed over two days. Practical considerations related to the school's schedule necessitated the use of unequal numbers of observations during each phase of the study.

Results

The mean percentage of time spent by *Ss* in the seven behavioral categories during baseline, contingency, and follow-up is presented in Table 1.

A least-squares analysis of variance with the effects of *Ss* and time absorbed (3) was performed on these data as indicated in Table 2. The analysis revealed a significant main effect for category ($F = 8.261$; $df = 6, 72$; $p < .0005$), indicating that student behavior was unevenly distributed across categories, as would be expected. Of greater importance, a significant category × time interaction ($F =$

3.996; $df = 12, 72$; $p < .01$) suggested that the distribution of students' behavior across categories changed significantly as a function of baseline, contingency, and follow-up.

Specific comparisons of category X time means using Tukey's procedure (8) demonstrated that academic cooperation with the teacher was significantly greater from baseline to contingency and from contingency to follow-up ($p < .05$). In addition, academic cooperation with students, the behavior of major interest in the present study, showed a substantial increase ($p < .10$) from baseline to contingency, and no significant decline from contingency to follow-up. Academic work alone decreased substantially ($p < .10$) from baseline to follow-up.

It can be seen from Table 1 that although changes in other behavioral categories over time failed to meet a rigid criterion of statistical significance, changes did occur in the appropriate directions. For example, it is particularly noteworthy that the three *Ss* who engaged most frequently in academically irrelevant behavior (other behavior) during baseline showed marked decreases for this behavioral category during contingency and follow-up.

Table 2.–ANOVA Summary Table for Mean Percentage of Time Spent by Subjects in Seven Behavioral Categories during Baseline, Contingency, and Follow-up[1]

SOURCE	*df*	*SS*	*MS*	*F*
Category	6	7881.962	1313.660	8.261*
Category X time	12	7625.867	635.489	3.996**
Error	72	11449.600	159.022	

*$p < .0005$
**$p < .01$

[1]The main effects of time and subjects have been absorbed, following Harvey (3).

Table 1.–Mean Percentage of Time Spent by Subject in Category

Ss	Academic Work Alone			Cooperative Academic Interaction with Student			Cooperative Academic Interaction with Teacher			Cooperative Non-Academic Interaction with Student		
	B	C	F	B	C	F	B	C	F	B	C	F
S_1	8	11	3	11	36	29	13	20	48	10	5	3
S_2	56	25	7	5	30	29	0	19	48	1	3	1
S_3	37	22	0	26	38	1	8	26	99	11	2	0
S_4	41	33	20	7	16	24	0	12	48	9	6	0
S_5	24	27	44	10	14	35	15	9	0	15	5	7

Ss	Cooperative Non-Academic Interaction with Teacher			Negative Interaction			Other Behaviors		
	B	C	F	B	C	F	B	C	F
S_1	4	10	7	0	1	3	54	17	7
S_2	4	8	0	6	3	0	28	12	15
S_3	8	4	0	1	0	0	9	8	0
S_4	4	8	0	7	2	5	32	23	3
S_5	11	11	1	4	13	5	21	21	8

B = Baseline
C = Contingency
F = Follow-up

Standard error for a category X time mean ($N = 5$) = 5.64
Standard error for a category mean ($N = 15$) = 3.25
Standard error for a time mean ($N = 35$) = 2.13

Discussion and Summary

The results of the present study suggest that peer interactions in a class of five emotionally disturbed children can be successfully manipulated in positive and academically relevant directions with the addition of peer-oriented contingencies to a more traditional ongoing token economy program. The strength of such manipulations is attested to by the fact that statistical significance was achieved or approached for relevant behavior changes, a more rigid criterion than is commonly applied to behavior modification research. Further, it is evident that behavioral changes, once established, persisted over time in the present study. It should be pointed out that both the classroom and the subject characteristics in the present study were relatively unique: Class size was relatively small, and the class members were probably more homogeneous with respect to race, socioeconomic status, IQ, and other factors than would be expected for many comparable special education classes, but the group was more heterogeneous with respect to age.

Results showed the efficacy of utilizing behavior modification principles in bringing about specific changes in the behaviors of a class of five emotionally disturbed children. It is clear from the positive outcome of this program that the use of behavior modification techniques to manipulate peer interaction, in addition to other target behaviors, is desirable. This technique appears to have considerable potential for use with students with behavior problems or other emotional difficulties since interaction is often a major problem for these children.

FOOTNOTE

1. We wish to acknowledge the generous assistance of Dr. Prentiss Shilling, Department of Experimental Statistics, Louisiana State University, and Mr. George S. McLean, Supervisor of Special Education, East Baton Rouge Parish School Board. We appreciate the encouragement of Dr. Donald L. Hoover, General Coordinator, East Baton Rouge Parish School Board.

REFERENCES

1. Buell, J.; Stoddard, P.; Harris, F.; and Baer, D., "Collateral Social Development Accompanying Reinforcement of Outdoor Play in a Preschool Child," *Journal of Applied Behavioral Analysis,* 1: 167 - 173, 1968.
2. Coleman, R., "An Economical Model of the Engineered Classroom," *Psychological Reports,* 28: 963 - 966, 1971.
3. Harvey, W., *Least-Squares Analysis of Data with Unequal Subclass Numbers,* United States Department of Agriculture, Agricultural Research Service, 20 - 8, July 1960.
4. Kirby, F.; and Toler, H., Jr., "Modification of Preschool Isolate Behavior: A Case Study," *Journal of Applied Behavioral Analysis,* 3: 309 - 314, 1970.
5. O'Leary, K. D.; and Becker, W., "Behavior Modification of an Adjustment Class: A Token Reinforcement Program," *Exceptional Children,* 33: 637 - 642, May 1967.
6. O'Leary, K. D.; and Drabman, R., "Token Reinforcement Programs in the Classroom: A Review," *Psychological Bulletin,* 75: 379 - 398, 1971.
7. Walker, H.; and Hops, H., "The Use of Group and Individual Reinforcement Contingencies in the Modification of Social Withdrawal," *Educational Research Information Center,* 6: 1 - 45, May 1972.
8. Winer, B. J., *Statistical Principles in Experimental Design,* McGraw-Hill, New York, 1962, pp. 77 - 89.
9. Zimmerman, E. H.; and Zimmerman, J., "The Alteration of Behavior in a Special Classroom Situation," *Journal of the Experimental Analysis of Behavior,* 5: 59 - 60, 1962.

THE EFFECTS OF FRUSTRATION ON THE FIGURAL CREATIVE THINKING OF FIFTH GRADE STUDENTS[1]

K. BRADLEY FROST
University of Georgia

ABSTRACT

In this study 24 fifth graders, 12 males and 12 females, were individually tested twice, once under frustrating conditions and once under non-frustrating conditions, in order to determine whether their creative expression would be affected by frustration. The frustrating condition consisted of being prematurely halted from completing a familiar task with the opportunity of receiving a reward upon its completion. The non-frustrating condition allowed the subject to finish the task and be rewarded. Changes in fluency, flexibility, originality, and elaboration as measured by the Torrance Tests of Creative Thinking, Figural Forms A and B, were investigated. On all components except originality, nonsignificantly higher scores were found under the frustrating conditions. Females scored higher than males on all four components. The only statistically significant interaction effect found was the sex X treatment interaction for elaboration.

ALTHOUGH THE VERY NATURE of our existence precludes the possibility of a frustration-free environment, the effects of frustration on mental functioning should be better understood so as to minimize its harmful influences and maximize its benefits. The scope of this exploratory study is to study the effects of frustration on the creative thinking of normal, healthy fifth grade children.

A considerable body of research (2) indicates that frustration and other stresses may cause either improvements or decrements in human performances depending

upon the intensity and duration of the frustration and the status of the person experiencing frustration. An earlier study by Turner (5) involving 55 emotionally disturbed children ranging in age from 6 to 18 years showed that even moderate frustration resulted in decrements in creative functioning as measured by the Torrance Tests of Creative Thinking, both figural and verbal. Many people believe, however, that mild frustration facilitates creative thinking and may even be necessary for it to occur.

The major purpose of the present study was to determine whether moderate frustration, such as that used by Turner, when applied to normal, healthy children would result in decrements of performance in figural creative thinking. Secondary purposes were to determine the interaction effects of level of academic achievement and sex with frustration on figural creative thinking. It was hypothesized that frustration would increase the variability of the performance of healthy fifth grade children but that it would not cause a decrement in performance. Further, it was hypothesized that level of academic achievement and sex would not influence the effects of frustration on figural creative thinking.

Procedures

The statistical design of the study is a simple repeated measures design with one within-subject variable (frustration) and two between-subject variables (sex and academic achievement level). The *S*s, in essence, acted as their own controls. Each *S* was tested twice, once under experimental conditions and once under control conditions.

The *S*s were drawn randomly from the fifth grade population of a rural intermediate school in Northeast Georgia. The *S*s were mostly from middle-class families. Eight *S*s, four male and four female, were selected randomly from each of the three achievement levels established in the school.

Experimental Treatment

The experimental procedures were as follows:

Step 1.–The *S*s were located by the researcher during the school day and escorted by him to a spare classroom where treatment was presented individually. During this time the researcher explained to the *S* that he had been chosen to help in an experiment for the University of Georgia if he so desired.

Step 2.–In the classroom was a table on top of which sat a large box filled with various types of candy bars. The *S* was seated by the box and given the following instructions: "Before we begin I want you to relax. This is not a test. I want to see how you can use your imagination. But before we do that, I want to give you a chance to win a candy bar. All you have to do is complete this word search puzzle in the time I allow and you may have your choice of any candy bar in the box. Are you ready? Go!"

At the end of two minutes, before the *S* could complete the puzzle but after he was totally involved, the researcher called time. At this point, the researcher closed the box of candy and removed it from the table.

Step 3.–Immediately following the frustrating experience, the Torrance Test of Creative Thinking (TTCT) was administered to each *S*. The standard directions provided in the test booklets were read to the *S*, with further explanation if requested by the *S*.

Step 4.–Approximately two weeks later the *S*s were located again and escorted to the same classroom.

Step 5.–Each *S* was presented with the box of candy bars and a similar word search puzzle to complete. This time the *S* was allowed to complete the puzzle and select his or her choice of candy.

Step 6.–Immediately following the non-frustrating experience, the TTCT was again administered to each *S* using the standard directions.

The sequence in which Steps 2 and 5 were administered to each *S* was randomly determined.

Word Search Puzzle

Two word search puzzles were used as a vehicle to induce frustration in the *S*s. Puzzles of this type are easily found in published books of crossword and other word puzzles. These two puzzles were constructed by one of the *S*s' teachers to insure the students' familiarity with the words. The object of the puzzle was to locate and circle all given words. This type of spelling exercise was familiar to all the *S*s since they were regularly assigned such a puzzle to complete. It is primarily because of the *S*s' knowledge of these puzzles and apparent ease in completing them that the researcher chose this instrument in hopes of eliciting true frustration by prematurely halting a familiar task. To determine the time limit to allow before halting the *S*s short of finishing each puzzle, the researcher worked each puzzle himself and achieved a minimum completion time of four minutes. A time limit of two minutes, or half the time needed for the researcher to finish each puzzle, was arbitrarily selected. This would hopefully allow enough time for the *S* to be fully involved in the task but not allow any *S* to finish.

Torrance Tests of Creative Thinking

For this study the TTCT Figural Forms A and B were used to measure creative expression. The researcher chose the Figural instead of the Verbal, or both, for the following reasons: Since the test was administered individually, the Figural took less time than the Verbal; the subjects have an art class each day, so figural expression was facilitated; reading inadequacies which could have been reflected in the Verbal (especially in the lower achievement group) were minimized by using the Figural form.

The two alternate forms of the Figural A and B were used to provide control against interaction of the tests

since the second testing period ranged from one to three weeks, with an average of two weeks, after the first. Figural Form A was used for the first testing situation and Figural Form B was used for the second. In the TTCT Norms-Technical Manual (4), Torrance provides evidence that the two forms are equivalent. Abundant evidence supporting the high validity and reliability of the TTCT can also be found in the TTCT Norms-Technical Manual.

Scoring

The tests were scored by the researcher, who obtained an interscorer reliability coefficient of better than .95 for each component with a trained scorer who is responsible for supervising scores for the TTCT on behalf of the Personnel Press Scoring Service.

The TTCT produces a composite score made up of four components: fluency, flexibility, originality, and elaboration. Fluency is the number of relevant responses. Flexibility is the number of different categories of responses and has the same maximum scores as fluency. Originality is based largely on the statistical infrequency of responses. Elaboration is the number of embellishments on the responses, with one point given for each idea communicated by each object, other than the minimum basic idea. There is no maximum score for elaboration.

To insure equivalency of the two forms of the TTCT, the raw scores were converted to standard or *t*-scores.

Results and Conclusions

The researcher computed a univariate ANOVA table for each of the four components to determine statistical significance set at the .05 level.

Although there was no statistically significant main effect, the results in Table 1 show that frustration seemed to increase creative behavior for all components except originality. The differences in the standard deviations of the experimental and control groups for fluency and flexibility may indicate an increase in variability of performance under frustrating conditions, which is also in accord with the researcher's hypothesis.

Results in Table 2 point out that the only statistically significant effect found was the sex X treatment interaction for elaboration. This may be interpreted as meaning that the effects of frustration on this component depend on whether the *S* is male or female. The difference between the experimental and control means, presented in Table 3, is 4.92 for males, whereas the difference for females is −5.08. The S X T interaction means point to the idea that males produce less elaboration under frustrating circumstances and that females produce more elaboration when frustrated. This interpretation seems compatible with our society's encouragement of "embroidering" behavior in females and discouragement of it in males. Also, the idea that under stress humans tend to behave in the manner familiar to them is in accord with this view.

The results for originality produced no significant difference; however, the S X T interaction did closely approach the .05 significance level.

The female *S*s scored higher than the males in all four components. This is contrary to Gallagher's report (1) that boys function better and achieve higher scores than girls on tasks requiring non-verbal performance. The higher female scores are also surprising in light of our society's traditional tendency to train males generally to produce more and better than females.

The academic achievement levels produced interesting data, with the average group scoring highest in all components except elaboration, where the high achievement group scored only 2.0 points better. The low achievement group scored the lowest in all four components.

So it seems that, contrary to the secondary hypothesis, sex does influence the effects of frustration on figural creative thinking in normal fifth grade children.

Although no definite conclusions can be made from this study, the most obvious implication may be seen as one of generating hypotheses involving the effects of frustration on creativity. One such hypothesis may involve the different effect that frustration seems to have on males and females. Since the females in this study were generally more creative, particularly under frustrating circumstances, a look into the possibility of academic environments containing different levels of frustration for males and females is plausible.

Another hypothesis might deal with the possibility that certain levels of frustration could be instrumental in increasing creative performance. Since a result of this study was a trend for higher creativity scores to be produced by induced frustration, further investigation may reveal the existence of such an optimal level of frustration which could possibly be utilized to maintain an optimal level of creativity in the classroom.

A third hypothesis could be formulated from the possibility that academic achievement level may not be a predictor of creative abilities. This idea is supported by the tendency of the average achievement level *S*s to demonstrate higher creativity scores than any of the other subjects. Therefore, using a measurement of creative abilities in addition to academic achievement level as a criterion for classifying or grouping students might prove to facilitate optimal educational development.

In summary, then, the major implication of this study seems to be one of stimulating hypotheses about the effects of frustration on creativity. However, the researcher recognizes that future investigations will be necessary in order for any of the above hypotheses to be substantiated. The researcher also recognizes the existence of certain limitations of this study, specifically, those limitations caused by the figural subtests of the TTCT since two alternate forms were used; those caused by this research design since the degree of frustration is not measured; those

Table 1.–Means and Standard Deviations for Fluency, Flexibility, Originality, and Elaboration by Treatment, Sex, and Academic Achievement Level

	Fluency		Flexibility		Originality		Elaboration	
	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*
Treatment:								
1. Frustration	42.75	10.64	44.21	10.01	60.50	19.59	57.33	11.72
2. Non-frustration	41.12	9.24	43.75	9.66	64.67	10.50	57.25	12.30
Sex:								
1. Male	41.42	12.22	43.29	10.74	60.95	21.14	56.38	10.66
2. Female	42.46	7.06	44.67	8.80	64.21	19.00	58.21	13.16
Academic achievement level:								
1. High	40.81	7.88	44.13	8.47	64.88	19.52	60.88	11.87
2. Average	46.38	12.86	48.06	10.44	69.81	22.45	58.81	12.00
3. Low	38.62	6.70	39.75	8.90	53.06	14.23	52.19	10.66

Table 2.–*F*-Ratios for Fluency, Flexibility, Originality, and Elaboration

Subject Variables	Fluency	Flexibility	Originality	Elaboration
Academic achievement level (A)	1.46	2.00	2.88	1.47
Sex(S)	0.07	0.16	0.31	0.18
Treatment(T)	1.01	0.06	0.76	0.00
A × S	0.69	1.86	1.23	0.70
A × T	0.22	0.38	2.80	0.15
S × T	1.35	0.06	3.87	4.56*
A × S × T	1.02	0.40	0.01	0.69

*$p < .05$ Tabled $F = 4.41$

Table 3.–Interaction Means of Treatment × Sex for Elaboration

	Treatment	
Sex	Frustration	Non-frustration
Male	53.92	58.83
Female	60.75	55.67

caused by the restricted population which resulted in such a small sample; and also those caused by the possibility of the Ss being affected by previous treatments.

FOOTNOTE

1. This paper is based on an M. A. thesis written under the direction of Dr. E. Paul Torrance, University of Georgia, 1973.

REFERENCES

1. Gallagher, J., "Productive Thinking," in M. Hoffman and L. Hoffman (eds.), *Review of Child Development Research: Vol. 1*, Russell Sage Foundation, New York, 1964.
2. Torrance, E. P., *Constructive Behavior: Stress, Personality and Mental Health*, Wadsworth Publishing Co., Belmont, Calif., 1965.
3. Torrance, E. P., *Torrance Tests of Creative Thinking: Directions Manual Scoring Guide (Figural Test Booklets A and B)*, Personnel Press, Lexington, Mass., 1972.
4. Torrance, E. P., *Torrance Tests of Creative Thinking: Norms-Technical Manual*, Personnel Press, Lexington, Mass., 1974.
5. Turner, S. M., *The Effects of Frustration, Sex, Mode of Behavior, and Organicity on Non-Verbal Creative Thinking of Emotionally Disturbed Children*, doctoral dissertation, University of Georgia, University Microfilms, Ann Arbor, Mich., No. 73-5798, 1972.

UNIQUE MULTIPLE LINEAR REGRESSION PROBLEMS FOR EACH STUDENT

GEORGE E. COUNTS
Southeast Missouri State University
Cape Girardeau, Missouri

ABSTRACT

The purpose of this report is to describe a process for creating unique multiple linear regression problems for each student. Three formulas were utilized to define intercorrelations between variables. A computer program was written based upon these mathematical relationships. To test the power of the program to "select" a random sample from variables with defined interrelationships, eleven examples were tested. On 33 sample correlation coefficients the expected number of significant correlations, 1.65, was larger than the number observed, 1. No difficulty is expected in preparing such problems for each student in a statistics or a research methods course, where professors may find problems of this type to be helpful.

IN AN EARLIER STUDY (1) an argument was made that unique statistical problems for each student in a statistics class have several advantages. Two types of problems were described. One type was for analysis with a *t*-test for the significance of the difference between independent means. The other kind was correlation and regression data which would allow inferences about a relationship between two variables.

Problems of these two types have been used by Southeast Missouri State University students for several years. Other types of problems have been simulated with limited effort and also utilized. Multiple correlation problems, however, proved to be both intriguing and aggravating. The objective is to provide each student with unique data which are random samples from a set of variables with a known multiple correlation coefficient. One approach is to control the relationships between all of the variables. In the process of maintaining relationships it is critical that means and standard deviations are also controlled. The resulting simulation problem is to define relationships between normally distributed variables in such a way as to provide an expected intercorrelation matrix as a limit when the sample size approaches infinity. The number of variables was arbitrarily limited to ten during this initial effort.

Procedure

The process of controlling interrelationships between variables depends upon use of appropriate formulas. Although some readers will be more interested in the results rather than how such results are achieved, an overview of the process may be helpful. Three basic formulas were finally used in the simulation process. The first formula is a standard score regression equation for predicting *z*-scores on variable *I* from variable *I*-1 through variable 1.

$$z_I' = \beta_{I\text{-}1}\, z_{I\text{-}1} + \beta_{I\text{-}2}\, z_{I\text{-}2} + \ldots + \beta_1\, z_1 \qquad (1)$$

The order of beta weights and *z*-scores here is different from the traditional (3:62). This order is chosen to fit the need to relate each new variable correctly to those variables which have already been included. The index *I* is at most ten and at least two.

The set of beta weights ($I > 2$) in each prediction equation (nine equations for ten variables) was calculated by calling a subroutine (MINV) provided in the Scientific Subroutine Package (SSP) for IBM 360 Model 40 users. The subroutine requires values from the appropriate (expected) intercorrelation matrix and returns values used to calculate the beta weights. This subroutine is one of a set which was written to provide a multiple linear regression analysis.

A second formula was used to determine the variance of the predicted values in Formula 1. The formula for the variance "of a composite of any number of weighted components" (2:421)

$$\sigma_{ws}^2 = \Sigma w_i^2 \sigma_i^2 + 2\Sigma r_{ij}\, w_i\, \sigma_i\, w_j\, \sigma_j$$

where $i < j$

can be simplified as long as the variables on the right side of the regression equation have standard deviations equal to one. In this context the standard deviations are all equal

to one because variables are in standard score form. This revised formula is stated below:

$$\sigma^2_{ws} = \Sigma w_i^2 + 2\Sigma r_{ij} w_i w_j \qquad (2)$$

where $i < j$

For this application, the weights (w_i and w_j) are beta weights. This variance represents the variance which can be predicted from other variables.

Finally, the amount of unique error variance (σ^2_e) for Variable I must be calculated. As the total variance for z-scores is one and predictable variance (σ^2_{ws} in Formula 2) is less than or equal to the total variance, error variance is equal to one minus predictable variance.

$$\sigma^2_e = 1 - \sigma^2_{ws} \qquad (3)$$

For $I = 2$ the beta value is the relationship between Variables 1 and 2 (ρ_{12}) and error variance is equal to one minus the square of the correlation coefficient ($\sigma^2_e = 1 - \rho^2_{12}$). For $I > 2$, beta weights and error variance are more difficult to determine, but, as indicated previously, the MINV subroutine provides beta weights, and Formulas 2 and 3 provide the unique error variance.

Given these formulas, and defining z_{e_i} as the ith random sample from a unit normal distribution ($\mu_{z_e} = 0$, $\sigma_{z_e} = 1$), the following formulas define interrelationships between variables:

$$z_1 = z_{e_1}$$

$$z_2 = \beta_{1_1} z_1 + \sqrt{1 - \beta^2_{1_1}}\, z_{e_2} \quad (\text{or } z_2 = \rho_{12} z_1 + \sqrt{1 - \rho^2_{12}}\, z_{e_2})$$

$$z_3 = \beta_{2_1} z_2 + \beta_{1_2} z_1 + \sqrt{1 - \sigma^2_{ws_1}}\, z_{e_3}$$

$$z_4 = \beta_{3_1} z_3 + \beta_{2_2} z_2 + \beta_{1_3} z_1 + \sqrt{1 - \sigma^2_{ws_2}}\, z_{e_4}$$

$$z_5 = \beta_{4_1} z_4 + \beta_{3_2} z_3 + \beta_{2_3} z_2 + \beta_{1_4} z_1 + \sqrt{1 - \sigma^2_{ws_3}}\, z_{e_5}$$

$$z_{10} = \beta_{9_1} z_9 + \beta_{8_2} z_8 + \beta_{7_3} z_7 + \beta_{6_4} z_6 + \beta_{5_5} z_5 + \beta_{4_6} z_4 + \beta_{3_7} z_3 + \beta_{2_8} z_2 + \beta_{1_9} z_1 + \sqrt{1 - \sigma^2_{ws_8}}\, z_{e_{10}}$$

The formulas above cover only the first five variables and Variable 10. The other four may be derived in the same manner. As the notation implies, the beta weight for a variable must be recalculated for each different equation in which the variable appears. Also, it should be noted that the z-score for Variable 1 is set to the first random "error." This is appropriate because all variance in Variable 1 is unique variance until other variables are generated. Also, z-score notation is appropriate for each variable. Formula 2, when applied to the weighted components, simplifies to $\sigma^2_{ws} + (1 - \sigma^2_{ws})$ or one.

Substantial time is required to incorporate these formulas into computer programs, to remove errors from each program, and to verify that output is acceptable.[1]

Results

In testing a computer program based upon the preceding formulas, eleven examples were drawn from Guilford's text (2:404) and are presented below in Table 1. An attempt was made to simulate sampling from populations with these interrelationships.

Table 1.–Correlations between Variables 1, 2, and 3 and the Multiple Correlation Coefficient for 11 Examples

EXAMPLE	ρ_{12}	ρ_{13}	ρ_{23}	$R_{1.23}$
1	.4	.4	.0	.57
2	.4	.4	.4	.48
3	.4	.4	.9	.41
4	.4	.2	.0	.45
5	.4	.2	.4	.40
6	.4	.2	.9	.54
7	.4	.0	.0	.40
8	.4	.0	.4	.44
9	.4	.0	.9	.92
10	.4	.2	−.4	.56
11	.4	−.4	−.4	.48

In each simulation trial 2000 values were drawn for each variable. The intention was to limit sampling error by this relatively large sample size. The GAUSS and RANDU subroutines (from the IBM 360 SSP) were used to allow sampling from normal distributions.

Table 2 contains statistical results for each example. The first three columns show sample means for each variable. All means are approximately zero as expected. The next three report observed sample standard deviations and are approximately equal to one. (The computer program also allows translation of z-scores into raw scores with specified means and/or standard deviations.) The final three columns are empirical estimates of the population values in Table 1. For each example the departures from expected values were judged to be sampling error.

To support this opinion, each member of each pair (sample value and corresponding parameter value) was subjected to Fischer's Z transformation (4:186). The transformed parameter value minus the transformed sample value was then divided by the standard error ($1/\sqrt{N-3} = \sqrt{1997} = .022$). The probability ($p$) of securing the

Table 2.—Means, Standard Deviations, and Pearson Correlation Coefficients for Variables 1, 2, and 3 by Example from Simulation Runs

EXAMPLE	$\overline{X}_1$	$\overline{X}_2$	$\overline{X}_3$	s_1	s_2	s_3	r_{12}	r_{13}	r_{23}
1	.051	−.001	.012	.990	1.000	.991	.4076	.4095	.0067
2	−.057	−.024	−.009	1.026	1.011	1.014	.4060	.3825	.4082
3	−.009	.020	.004	1.006	.999	1.014	.4069	.3997	.8979
4	.023	.017	−.009	.988	.957	.972	.3660	.1871	−.0259
5	.016	.010	−.024	1.003	1.040	1.008	.4490	.2193	.4217
6	.012	−.015	−.005	1.016	.990	.981	.4051	.1923	.8982
7	.033	.054	.006	1.002	1.004	.981	.4049	.0219	.0115
8	−.001	−.002	.015	1.025	1.010	.988	.4134	.0140	.4168
9	.045	.029	.015	1.017	.990	.976	.4234	.0168	.8959
10	−.011	.033	.007	1.039	.955	1.011	.3978	.2344	−.3949
11	−.003	.018	−.021	1.004	1.027	1.003	.4131	−.4027	−.4014

largest difference (Example 5, $r_{12} = .449$, $\rho_{12} = .400$) or a more extreme difference in either direction by chance is .007. The second most unlikely value (Example 4) was $r_{12} = -.366$ ($p = .072$). For the 33 sample values, only one was classified as more improbable than alpha levels of .05 or .01.

Although the most extreme case (considered in isolation) seems rather improbable, it is the only one which demonstrates significant departure from expected sampling error. In a group of 33 such tests, we should expect one or more—(.05) (33) or 1.65—such results. There is no clear evidence that the set of sample correlations is significantly different from expected values. This process could be repeated with an even larger sample size if necessary.

The next problem is to relate these results to the multiple correlational setting. The following correlation formula (2:404) could have been used for each of the eleven examples in Table 2:

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23}}{1 - r^2_{23}}$$

However, as indicated previously and as implied by this formula, the multiple correlation coefficient is completely dependent upon the relationships between Variables 1, 2, and 3. If each sample value (r_{12}, r_{13}, and r_{23}) is approaching the corresponding parameter value (ρ_{12}, ρ_{13}, and ρ_{23}) as a limit, then the squared sample multiple correlation coefficient ($R^2_{1.23}$) also is approaching its limit.

Finally, an additional test was made to determine the computer program response to an intercorrelation which is impossible. For each member of the set (ρ_{12}, ρ_{13}, ρ_{23}) a value of −1.00 was assigned. The matrix was identified as being singular by the MINV subroutine, and no correlations were attempted.

Conclusions

The following conclusions seem to be valid:

1. The flexibility of the program is substantial; no difficulty is expected in the generation of statistics problem for students.

2. The experience of computer simulation of such variables may provide additional appreciation for the complexity and simplicity of the statistical method. The development of a computer program using these formulas is not too difficult. Advanced students might prefer to start from the problem definition stage.

3. In some situations it may not be necessary to know the true or expected relationships. If interval-scaled scores from a normal distribution are converted to scales of less precision (and theory does not provide exact relationships) a comparable "sample" with an even larger "sample" size can be used as a population. This population may or may not include the student's sample.

4. With limited effort, the process could be modified to provide sampling models which are alternatives to sampling from unrelated variables—those for which all beta weights are zero. Given an hypothesized intercorrelation matrix, empirical estimation of the probability of securing a multiple correlation coefficient less than (or greater than) the observed coefficient is possible. Rejection of the null hypothesis at an improbable value plus failure to reject hypothesized values at a highly probable value (using simulation methods) are closer to the real objective. Simulation trials, however, may be impractical with large samples.

5. In addition to the potential use in teaching statistics courses which include multiple linear regression techniques, this process could also be used to generate "data" for analysis by a student in a research methods course. At least some courses of this type provide little or no opportunity for the student to demonstrate or improve his competence in analyzing such data and reporting his findings

and conclusions. Dissertation committee members might be reassured by the knowledge that a particular research plan, based upon multiple linear regression, was carried out to its logical conclusion, that the results were well organized and accurate, that conclusions drawn were reasonable, and that the nature of the "population" has been shared with the investigator.

FOOTNOTE

1. Program listings for each program used in this process may be obtained from the author, Dr. George E. Counts, Southeast Missouri State University, Cape Girardeau, Missouri, 63701.

REFERENCES

1. Counts, G. E., "The Creation and Solution of Unique Statistical Problems for Each Student," *Journal of Experimental Education,* 37: 17 - 20, Spring 1969.
2. Guilford, J. P., *Fundamental Statistics for Psychology and Education,* McGraw-Hill, New York, 1965.
3. Kerlinger, F. N.; and Pedhazur, E. J., *Multiple Regression in Behavioral Research,* Holt, Rinehart and Winston, New York, 1973.
4. Klugh, H. E., *Statistics: The Essentials for Research,* John Wiley, New York, 1974.

PRESCHOOL INFLUENCES ON OCCUPATIONAL KNOWLEDGE OF SEVEN-YEAR-OLDS: A PROSPECTIVE STUDY[1]

THOMAS E. JORDAN
University of Missouri at St. Louis

ABSTRACT

A prospective study of 180 children from birth to age seven is reported. The criteria at age seven was knowledge of occupations as indicated by a pictorial test. The predictors were: seven aspects of the child and home at birth's, maternal IQ as tested when child was three years, a quantified description of the potential stimulating characteristics of the home at four years; and a measure of parental attitudes to schooling when child was age five. The data were subjected to a multivariate regression analysis. Social class data were the prime source of criterion variance.

THE PROBLEM STUDIED in this investigation is a description of the contribution of home and family variables to the degree and type of knowledge of occupations held by children in first grade. The importance of such information to people working on curricular aspects of vocational choice in the elementary school is considerable. Curricula to increase the quality of career choices by young persons need to exert an influence long before the years of adolescence. Presumably, career choice should begin, like other aspects of development schools wish to encourage, in the elementary school.

The preceding remarks are cast in the context of traditional views of career development and acquisition of occupational information. From a body of research and writing (1, 2, 19, 21), it is clear that school counselors' effectiveness in vocational counseling is greatly influenced by antecedent concepts of work. These concepts are forme well before the time for vocational selection emerges. Although information about formation of work concepts is incomplete (1, 13), the work of Gribbons & Lohnes (11) so indicates that children in the eighth grade have attained vocational self-concepts which show a good deal of stabilit in subsequent years. Recent work on the problem by Wehrly (22) indicates that early formation of work concepts is not related to parental occupation, as common sense might indicate. The problem is more subtle, and the

influences are not clear. From this it can be concluded that there is a need to explore the complex of early influences which determine knowledge of occupations.

However, there is a rather different intellectual framework within which the matter of origin of work concepts and occupational choice has appeared in the last few years. The problem has been most boldly presented by Jencks in his 1972 book, *Inequality: A Reassessment of the Effect of Family and Schooling in America* (12). Jencks observes that the social problems of inequality of attainment and status in our society are tied to schooling–an observation this writer has documented recently (13), and which has been examined in great detail by Duncan et al. (7). Jencks goes on to set forth a conclusion based on his analysis of the 1966 Coleman Report data: that simply raising the level of funding for schools, the concept of general support implicit in most states' *Foundation* program of assistance, will not do. We are thus left with the need to emphasize preparation for vocational choice on a rational, curricular basis. This is, of course, a view traditionally held by vocational experts, but it is one now articulated at the highest levels of educational strategy. To this discourse is added another point made by Jencks (12: 256):

> Our research suggests, however, that the character of a school's output depends largely on a single input, namely the characteristics of the entering children. Everything else–the school budget, its policies, the characteristics of the teachers–is either secondary of completely irrelevant.

It is this central matter of ". . . characteristics of the entering children. . ." and their explicit knowledge of arrangement of work occupations that this inquiry has pursued. The inquiry accordingly addresses the twin problems of describing occupational knowledge and drawing inferences for the development of curricular materials to develop career information.

Method

General Design

The general design is a prospective longitudinal inquiry with a multivariate analysis of data. Such an inquiry has been under way on a population of 1,000 newborns delivered in five St. Louis hospitals in the winter of 1966-67. The 1966-67 cohort has been studied in two portions: the winter group selected for this study refers to children studied on or very close to the seventh anniversary of their births, as opposed to a summer group of children studied annually, but six months after their birthdays.

The children in question were traced through an annual process of confirming addresses. Appointments were made to test the children individually in their homes using trained and experienced examiners matched by race. The number of children traced and tested was 284.

Variables

The criterion variable administered at age seven years was Fulton's *Test of Career Knowledge* (9). This is a 30-item picture recognition test composed of items from the categories of the *Dictionary of Occupational Titles* (DOT) (6). Specifically, the categories are: DOT # 1–professional, technical, and management occupations (9 items); DOT # 2–clerical and sales occupations; and DOT # 8–structural work occupations (5 items). A fourth category, miscellaneous occupations (5 items), is not reported here due to its heterogeneity. The test has an internal consistency (reliability) of .86, and validity, according to Fulton et al. (10), is demonstrated by conformity to three aspects of curricular validity, plus conformity to elements of the DOT.

The predictor variables were selected from a set of measures previously gathered during annual testing in an attempt to build a picture of family characteristics at specific time points in the preschool years. They were chosen to shed light on the influences determining career knowledge in first grade with a view to drawing inferences for curriculum development. The predictor and criterion variables are given in Table 1, and the following elements are those which are not self-explanatory:

Intelligence score is the raw score attained by the study child's mother on the Ammons' (1) Quick Test, (QT), a valid and reliable vocabulary-type instrument. This test was chosen because it draws on verbal skills relevant to the overall purposes of the prospective study, and because it could be administered under adverse home circumstances more easily than most tests of ability.

Maternal education level is a score from 1 to 5 which is assigned according to years of schooling. It ranges from a low score of 1 for elementary education only to a score of 5 for college education.

Paternal education level is a score from 1 to 7 which is assigned according to level of schooling. It ranges from a low score of 1 for elementary education to 7 for a graduate degree.

PATE *score* is the score on Medinnus' *Parent Attitude to Education* scale (18).

STIM *score* refers to a quantified description of the potential stimulating characteristics of the home, as developed by Caldwell (5).

SES *score* is a weighted three-factor description of the socioeconomic level of the home based on the breadwinner's level of schooling, occupational title, and level of income, as developed by McGuire and White (17).

Statistical Design

The analytic technique applied to the data was intended to (*1*) identify salient variables by means of the specific contribution to criterion variance, and (*2*) examine interactions of the variables identified as influential. The multi-

variate technique employed, the AID-4 interaction regression approach, chooses elements of a predictor series which meet predetermined criteria for contributing to criterion variance. Through a series of heuristic splits of predictors, a set of consecutive splits is contrived and extended. The splits or branching into subordinate but significant predictors creates a tree-like array of predictors in which both primacy of contribution to criterion variance and interactions may be identified (16, 20).

Results

Descriptive Findings

The data analyzed in this report consist of information from the developmental history of each study child from birth to age seven years. Complete information on all ten predictor variables plus criterion variables is reported for 180 *Ss* in Table 1. There it will be seen that the Fulton test criterion score has been additionally treated to provide three additional criterion scores: DOT subgroup scores for technical, clerical, and structural occupations, bringing the total number of criteria reported and analyzed to four.

It can be seen from Table 1 that the *Ss* are approximately balanced by sex (47% male). The racial composition of the study group is 20% black, which approximates the national proportion. The mean social class (SES) score of 51.59 in Table 1 is quite close to the mean of 59 (σ = 16) observed in 1966-67 for the birth cohort (15). The difference of 2 points is in the direction of a slightly higher social level. This insignificant trend is explained by the problem of tracing and testing children from the lowest social strata. This particular problem also explains why the inferential study group described here is less than the 280 cases tested. The children omitted are those for whom there was incomplete information on any of the ten predictor variables gathered during the preceding seven-year period. The mean level of schooling score attained by mothers is 2.93, which is a little less than four years of high school, on the average. The intelligence score mean reported for mothers in Table 1 translates into an IQ of 92. Paternal education level is only slightly higher than maternal; the mean of 4.19 in Table 1 indicates an average level of schooling just a little beyond high school for fathers. The birth order mean value is 2.88, indicating that the study children tended to be the third-born, having two older brothers and sisters. This fact is consistent with the information in Table 1 on the mean age at delivery of the mothers. The average mother was 26 years and four months at the time the child under study was delivered and enrolled in the longitudinal study. Again, on the average, she took the PATE scale at age 31 years (child age 5 years), and had a mean score of 59.02. This score is also similar to that observed for several hundred mothers on the average.

Regression Models

The basic technique of inferential analysis, the AID-4 method, employs mathematical regression of variables on a

Table 1.—Description of *Ss* Used in Multivariate Analyses (N = 180)

CHILD AGE	PREDICTOR/CRITERION	$\bar{X}$	σ
Delivery	Three-factor SES score [McGuire & White (17)]	51.59	15.77
Delivery	Race (%W)	80	
Delivery	Paternal education	4.19	1.50
Delivery	Maternal education	2.93	.98
Delivery	Birth order		
Delivery	Maternal age	26.31	6.35
Delivery	Sex (%M)	47	
3 years	Maternal QT raw score [Ammons & Ammons (1)]	38.94	5.12
4 years	STIM score [Caldwell (5)]	34.66	4.53
5 years	PATE score [Medinnus (18)]	59.02	7.01
7 years	Fulton test total score	22.68	3.03
7 years	Technical occupations score (DOT #1)	3.48	.98
7 years	Clerical occupations score (DOT #2)	4.25	.86
7 years	Construction occupations score (DOT #8)	3.72	.91

Table 2.–Regression Models of Fulton Test Criteria and Their Levels of Statistical Significance

CRITERION	R^2	df_1	df_2	F	p
Total score	.24	7	172	7.79	<.01
DOT technical group #1	.13	7	172	3.79	<.05
DOT clerical & sales group #2	.24	7	172	8.29	<.01
DOT construction group #8	.18	10	169	5.23	<.01

criterion. It is helpful to observe the materials listed in Table 2. There, the final regression models are given as described by the interaction/regression program. It may be seen that the R^2 values are mostly from .18 to .24, with the exception of the second criterion, for which the R^2 value is .13. While these values are not high, they are quite typical of values observed in other analyses at school entry age (15). In three of these cases the complex models listed are statistically significant at the .01 level of confidence, while the others are significant at the .05 and .03 levels. The five models generated are, accordingly, adequate for purposes of further analysis. A more detailed examination of the interaction patterns of "trees" generated through these regression models is now presented.

In Figures 1 - 4 the results of applying the interaction regression technique to scores obtained by children at age 78 months on Fulton's Test of Occupational Knowledge are seen. The four figures represent formulation of the scores. First, there is a total score, followed by Figures 2 - 4 representing subscores for three occupational subgroups–professional and technical, clerical and sales, and structural work–corresponding, respectively, to DOT categories # 1, 2, and 8.

In Figure 1, the full score on the Fulton test, five of the nine elements in the predictor set have been retained as significant factors. They are, in order, the three-factor social class score; race; birth order; maternal raw score on the Ammons' Quick Test of verbal intelligence; and paternal level of schooling. The last elements in the tree, groups 8 - 9, 12 - 13, and 14 - 15, were composed from data on maternal QT scores, paternal level of schooling, and birth order. These last three splits raised the proportion of assigned criterion variance, but not a a statistically significant level. Of the total variance accounted for by the full model, R^2 = .24, nearly two-thirds were explained by the first two factors, SES and SES-race. The series of splits which generated the model schematized in Figure 1 was generated through Group 3, whose levels 3 - 8 represent all but the high levels of socioeconomic status.

Figure 2 presents essentially the opposite process of elaboration through low levels of the prime variable. In this case the predictor is maternal age at the time of delivery of the child. The upper levels of delivery age in Group 3 were slightly extended late in the splitting process by Groups 12 and 13. In contrast, the Group 2 representing the younger mothers created the remaining ten cells of the tree. In this, the more extensive branch of the AID-4 tree, there is repetition of the role of mothers' PATE scores in Groups 8 - 9, and, later in the sequence of decreasing contribution to the regression model, in Groups 14 - 15. This last split raised the proportion of assigned variance to R^2 = .13. There is an interesting aspect to the predictor set schematized in Figure 2; it is that the first three predictors are all maternal traits–age at delivery of the study child, attitude to education, and attained level of education. Actually, all but one of the seven predictors is a maternal trait. The sole exception is the relatively uninfluential element of paternal level of education.

In Figure 3, the AID-4 diagram for the third criterion, the clerical and sales subgroup score, shows a tree which is asymmetric. Group 2 containing 40 children represents low scores (high levels) of social class scores. This branch was relatively unproductive, and was completed quite late in the process of raising R^2 values by maternal educational level in Groups 14 - 15. A majority of the children were in Group 3 and had lower levels of social class background; the AID-4 diagram was largely developed through them. The first split from Group 3 was by maternal QT score, and is seen in Groups 4 - 5. The lower branch ended in Groups 8 - 9, based on fathers' educational attainment. The mean of the lowest group of scores is in Group 8. The mean criterion score of Group 8 was 2.45, which is approximately 1.25 standard deviations below the mean given in Group 1 for N = 180 children. The more elaborated development of the AID-4 tree begins with the score for QT in Group 5. It splits into Groups 6 - 7 based on the STIM score, followed via Group 6 by the PATE score which expresses maternal attitude to education. The final split of this branch ends with Groups 12 - 13, which are based on maternal QT scores.

Figure 4 shows the selection, priority, and nature of the variables associated with the Fulton test score for the construction subgroup, occupational category #8 of the DOT. Quite the most complex tree, it represents a regression model which achieved an R^2 value of .23,

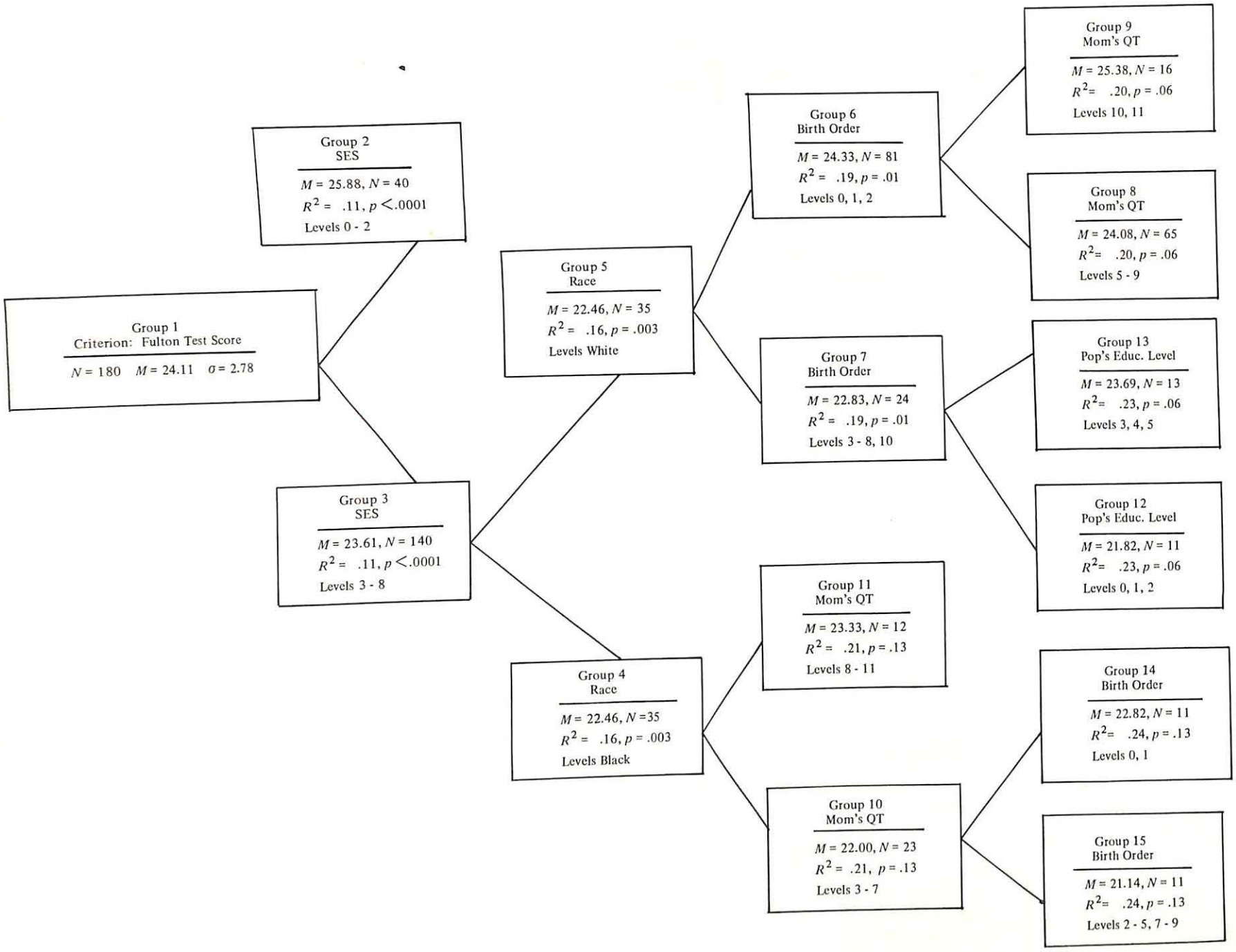


Figure 1.–AID-4 Tree for Total Score on Fulton Test of Career Knowledge

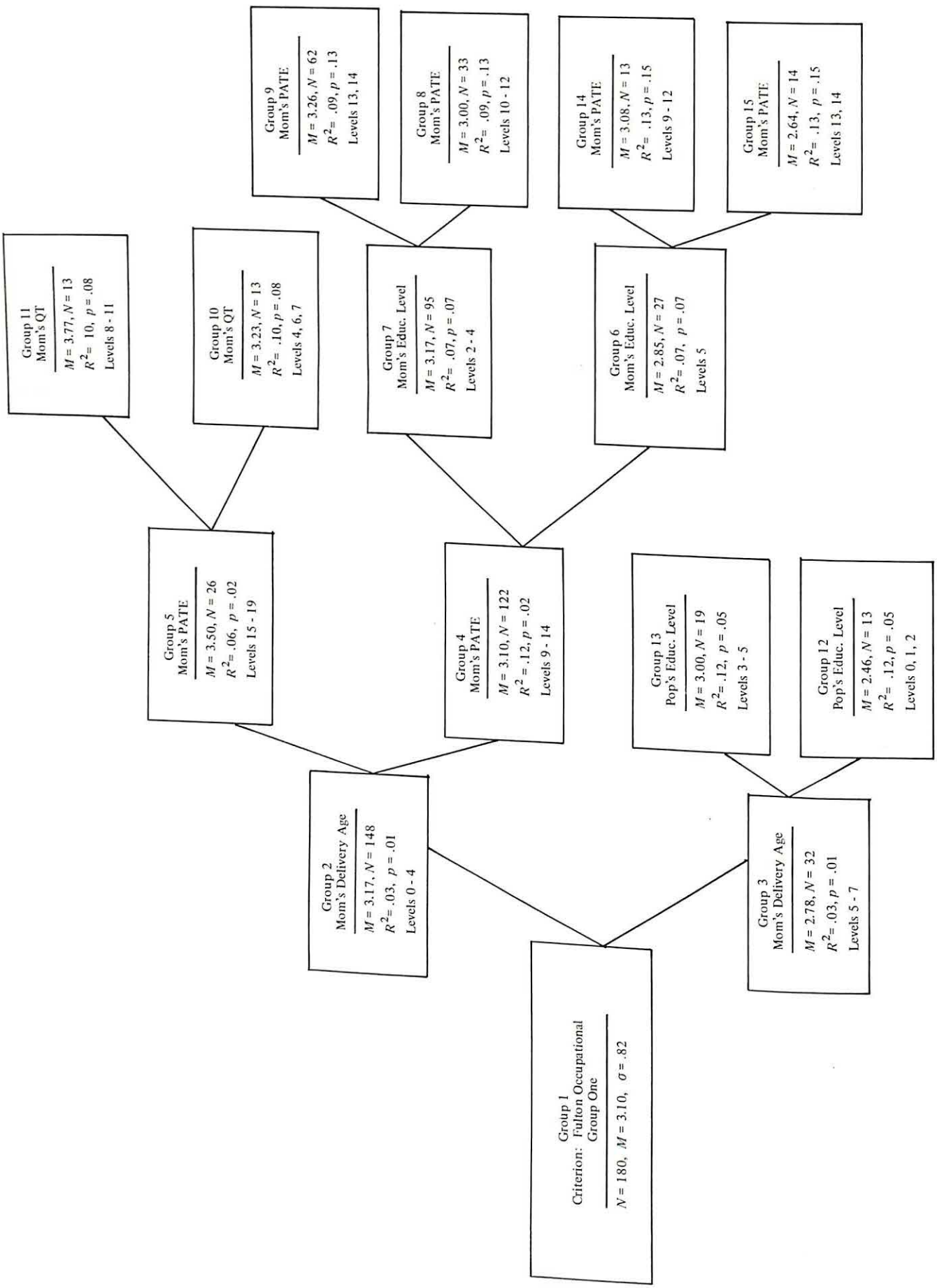


Figure 2.–AID-4 Tree for Professional-Technical Occupational Subgroup

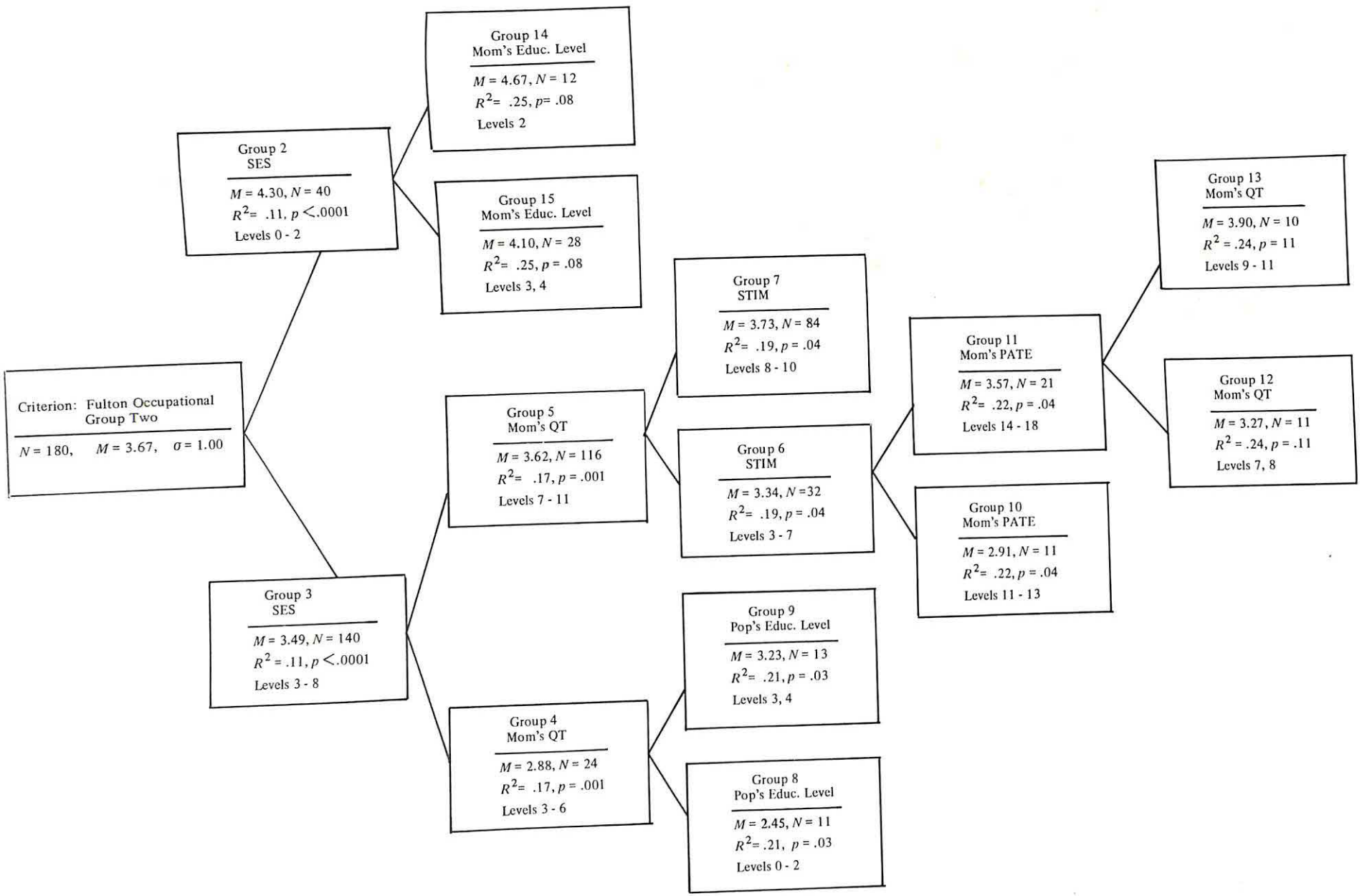


Figure 3.–AID-4 Tree for Clerical Occupational Subgroup

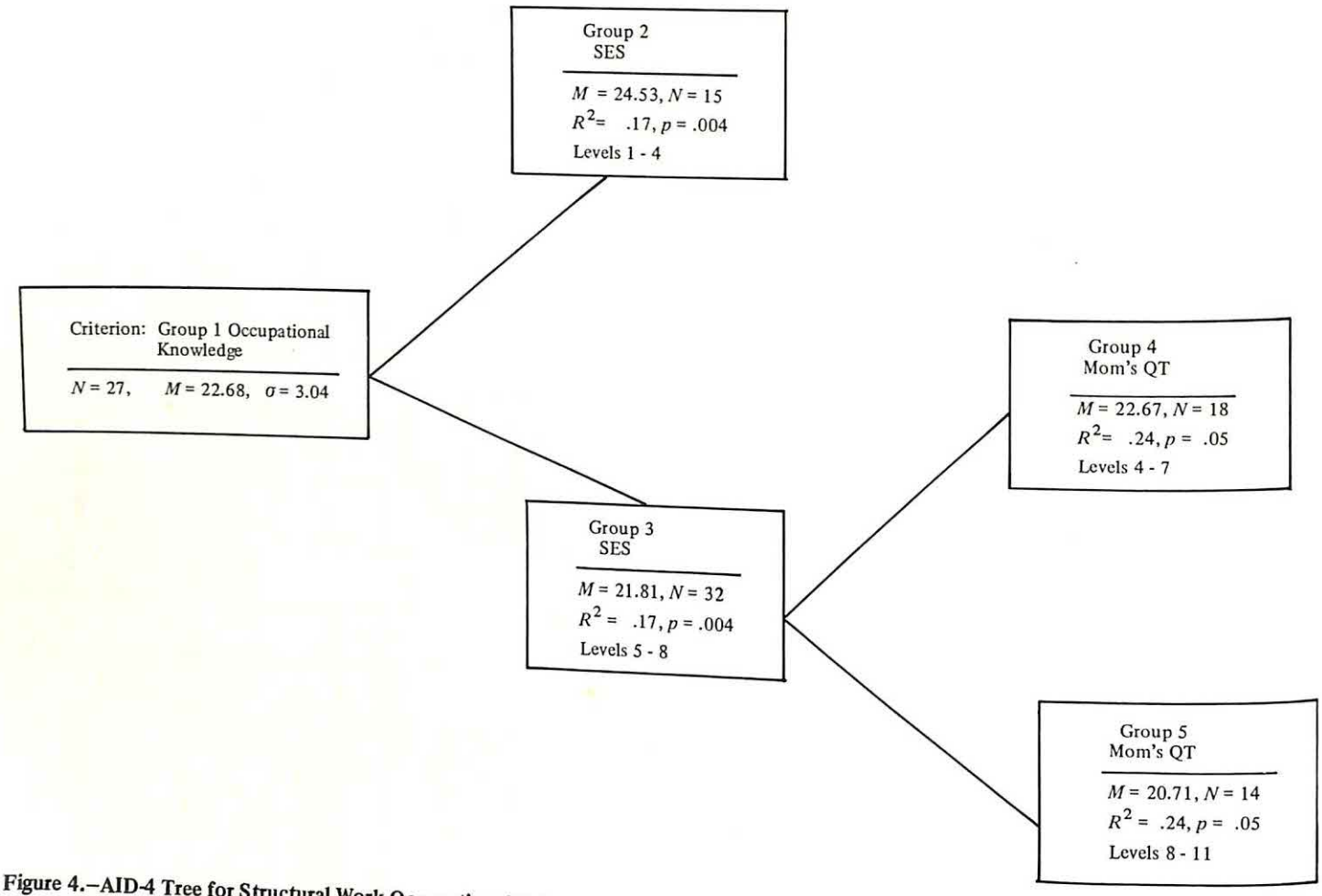


Figure 4.–AID-4 Tree for Structural Work Occupational Subgroup

through nine splits into 21 cells. The prime source of attributed variance is the three-factor social class score. The branch containing low levels of the SES score (*high* SES) is completed by low-numbered Groups 18 - 19 based on STIM score. The low side, Group 3, is the basis for the elaboration of the tree into 21 groups. Maternal age at delivery, Groups 4 - 5, is the variable through which the process of splitting began. On the high side of this variable, Group 4, subsequent elaboration through birth order, and delivery age, and, reciprocally, the STIM and PATE variables in Groups 16 - 17, and 20 - 21, can be seen. On the low side of delivery age, Group 5, elaboration through the variables STIM, maternal educational level, and sex is seen. This last finding is interesting because Groups 14 - 15 are the only instance of a boy/girl contribution in all five AID-4 interaction regression analyses.

Discussion

Findings

The major descriptive finding of the study is that there are interesting differences in the amount of occupational knowledge in the four groups examined: Based on the full score on the Fulton test, the average child's knowledge of all occupations presented by the criterion test is 80%. For the first occupational subgroup (DOT # 1), the mean number of occupations known is 34%. For the second occupational subgroup (DOT # 2), the mean score is 60%. For the third occupational subgroup (DOT # 8), the mean score is 86%. When reviewed by occupational categories the highest level of knowledge possessed by the children was for the construction occupations–86%, followed by clerical occupations–60%. Least knowledge is demonstrated for the professional and technical jobs–34%.

Turning to the AID-4 analyses, it is helpful to note the extent to which the groups' means in the trees represent a spreading-out of criterion scores. Table 3 gives the grand means and the highest and lowest group means from the four AID-4 analyses, wherein ± one *SD* values for the total scores on Fulton test are 26.89 and 21.33. The highest group projected by the AID-4 analysis in Figure 1 is within the boundary, but the low group is below the minus and sigma value. In the case of Figure 2, the high and low boundaries are 3.92 and 2.28. The high and low cells of the tree in Figure 2 approach but do not exceed the ± one sigma range. In the case of Figure 3, the boundary values

Table 3.–Grand Means and High Group and Low Group Means

ANALYSIS	MEAN	SIGMA	GROUP	HIGH MEAN	GROUP	LOW MEAN
Figure 1	24.11	2.78	2	25.88	15	21.25
Figure 2	3.10	.82	11	3.77	12	2.46
Figure 3	3.67	1.00	14	4.67	18	2.45
Figure 4	4.32	.79	18	4.92	6	3.70

Table 4.–Primary, Secondary, and Tertiary Predictors in AID-4 Regression Models

PREDICTOR	TOTAL SCORE	SUBSCORE ONE	SUBSCORE TWO	SUBSCORE THREE	(f)
Three-factor social class score [McGuire & White (17)]	1*		1	1	(3)
Race (%W)	2				(1)
Paternal education					
Maternal education		3			(1)
Birth order	3				(1)
Maternal age	1		2		(2)
Sex (%M)					
Maternal IQ [Ammons & Ammons (1)]			2		
STIM score [Caldwell (5)]			3	3	(2)
PATE score [Meddinus (18)]	2				(1)

*1 - ordinal position in AID-4 tree

are 4.67 and 2.67. The high and low groups have mean scores at or beyond these boundaries. For Figure 4 the boundaries are 5.11 and 3.53, and the high and low group means approach but do not exceed these values. In general, the spread of criterion scores from highest to lowest groups is almost from one *SD* above the mean to one *SD* below it. The AID-4 trees are satisfactorily spread out on either side of the grand mean scores.

Turning now to a consideration of the variables in the children's backgrounds which account for criterion variance of the four Fulton test scores, Table 4 lists the first three predictors in order of magnitude for four criterion scores; e. g., race is found once, in the tree for total score, as the second most influential predictor. Two of the predictors were not used in the first three splits in creation of the four AID-4 trees: the sex of the child and the level of education achieved by the child's father. Two variables used only once were maternal traits: educational level and attitudes to education (PATE). In contrast, one variable was used three times in the projection of AID-4 trees, and in each instance it was used as the first split in the three, the prime source of criterion variance. This variable is the three-factor social class score (SES), and it was the prime variable in Figures 3 and 4 and in the tree for the total score on the criterion Fulton scale in Figure 1.

On the basis of the finding that SES has such a powerful effect–as opposed, for instance, to the lack of an effect due to sex–it is helpful to look at Table 5. It shows predictor and criterion scores displayed by quartiles of the SES score; e. g., four mean Fulton test criterion scores are presented for four levels of the perinatal social class score. Inspection of both predictor and criterion scores in Table 5 shows the trend to higher scores with rising SES level.

Implications

In considering the implications of the preceding materials for construction of elementary school curriculum materials on occupations, it is first observed that there is an uneven degree of knowledge among seven-year-olds concerning the four occupational categories tapped by the Fulton

Table 5.–Mean Predictor and Criterion Scores Arranged by Social Class Score Quartiles

PREDICTOR/CRITERION	ALL *Ss* (*N* = 180)	(Low SES) FIRST QUARTILE (*N* = 37)	SECOND QUARTILE (*N* = 52)	THIRD QUARTILE (*N* = 43)	(High SES) FOURTH QUARTILE (*N* = 49)
Sex (%M)	47	32	50	53	51
Race (%B)	20	48	30	4	0
SES	51.59	70.75	60.26	49.53	29.71
Delivery age	26.31	26.18	24.25	26.23	28.67
QT (IQ)	92	84	90	92	100
Birth order	2.88	3.62	2.61	2.58	2.87
Mother's education	2.93	2.13	2.71	2.83	3.85
STIM	34.66	31.43	34.36	35.41	36.75
PATE	59.02	62.13	60.30	55.83	58.12
Father's education	4.19	2.70	3.55	4.23	5.95
Fulton test total score	24.03	22.29	23.73	24.11	25.59
Technical occupations score (DOT #1)	3.09	3.05	3.09	3.00	3.20
Clerical occupations score (DOT #2)	3.66	3.27	3.42	3.74	4.14
Construction occupations score (DOT #8)	4.30	4.10	4.30	4.13	4.59

Test of Career Knowledge. Seven-year-olds seem to be most informed about the construction occupations. In contrast, children seem to know less about professional and technical jobs. This suggests that small children begin school with a quite restricted sense of work opportunities, and points to a hiatus in their thinking which curricular materials could seek to remedy over a period of years, especially in relation to technical and professional challenges.

Within this imbalance across occupational groups there is the consistent influence of social class. Children from the higher social classes are more informed about all four occupational groups than children who are less favored. Interestingly, this bias does not raise knowledge of technical jobs disproportionately for the highest social class group. Rather, the bias is for greater knowledge of all occupations including those at variance with the child's personal background.

From the data of this inquiry we may identify some leads for strategy in developing curricular materials, although the low R^2 values of the regression models provide no mandate for radical innovation at this stage.

First, the high level of knowledge of construction jobs may reflect the tangible nature of such jobs. This suggests that concrete, experiential materials may be of prime value in broadening the range of jobs known to children.

Second, the absence of a sex effect in the set of major sources of criterion variance eliminates an initial sex discrimination factor in children's knowledge of jobs. Accordingly, it seems that boys and girls begin elementary school with essentially sex-less orientations to job knowledge. In view of the contrary pattern which emerges with the socialization process, it is helpful to know that the sex bias found in adolescent- and adult-women's job information is not innate or unavoidable.

Third, the absence of a race effect among the prime sources of criterion variance is much like the sex factor. That is, the limited occupational information of black, inner-city youth can be viewed as a developmental bias not evident at age seven years. Accordingly, the problems of job choice–and job accessibility–in adolescence may be viewed as developmental in both black youth and in their peers in the white community.

FOOTNOTE

1. This study was supported by grants from CEMREL, Inc., the National Institute of Education, and the state of Missouri.

REFERENCES

1. Ammons, C.; and Ammons, R., *The Quick Test,* Psychological Test Specialists, Missoula, Mont., 1962.
2. Barrow, H., "Development of Occupational Motives and Roles," in L. Hoffman and M. Hoffman (eds.), *Revies of Child Development Research,* Russell Sage Foundation, 1966.
3. Becker, R. L., "Vocational Choice: An Inventory Approach," *Educ. Train. Ment. Ret.,* 8: 128 - 136, 1972.
4. Brickell, H. M.; and Aslanian, C. B., *Attitudes towards Career Education,* Educational Testing Service, Princeton, 1972.
5. Caldwell, B. M., *The STIM Scale,* University of Arkansas, 1970.
6. *Dictionary of Occupational Titles,* 3rd ed., U. S. Employment Service, 1965.
7. Duncan, O. D.; Featherman, D. L.; and Duncan, B., *Socioeconomic Background and Achievement,* Seminar Press, New York, 1973.
8. Fulton, B., Vocational Development of Children, unpublished doctoral dissertation, University of Missouri at Columbia, 1971.
9. Fulton, B., *Fulton Test of Career Knowledge,* Evaluative Research Associates, Inc., St. Louis, 1973.
10. Fulton, B. J.; Marshall, J. C.; and Sokol, A. P., *Career Education Strategies,* Evaluative Research Associates, St. Louis, 1975.
11. Gribbons, W. D.; Lohnes, P. D., "Predicting Five Years of Development in Adolescents from Readiness for Vocational Planning Scales," *Journal of Educational Psychology,* 56: 244 - 253, 1965.
12. Jencks, C., et. al., *Inequality: A Reassessment of the Effect of Family and Schooling in America.* Colophon Books, 1973.
13. Jepsen, D. A.; and Dilley, J. S., "Vocational Decision-making Models: A Review and Comparative Analysis," *Review of Educational Research,* Chicago, 44: 331 - 349, 1974.
14. Jordan, T. E., *America's Children,* Rand-McNally, 1973.
15. Jordan, T. E., *The Natural History of 1008 Infants in the Readiness Years: Final Report,* National Institute of Education, Washington, D. C., 1974.
16. Koplyay, J., "Automatic Interaction Detection," *Multiple Linear Regression Viewpoints,* 3: 25 - 38, 1972.
17. McGuire, C.; and White, G., "The Measurement of Social Status," *Research Papers in Human Development,* University of Texas, 1955.
18. Medinnus, G. R., "Development of a Parent Attitude toward Education Scale," *Journal of Educational Research,* 56: 100 - 103, 1962.
19. Nelson, R. C., "Knowledge and Interests Concerning Sixteen Occupations among Elementary and Secondary Students," *Educational and Psychological Measurements,* 23: 741 - 754, 1963.
20. Sonquist, J. A.; Baher, E. L.; and Morgan, J. W., *Searching for Structure,* Institute for Social Research, University of Michigan, 1974.
21. Steward, L. H., "Relationship of Socioeconomic Status to Children's Occupational Attitudes and Interests," *Journal of Genetic Psychology,* 95: 111 - 136, 1959.
22. Wehrly, B. L., "Children's Occupational Knowledge," *Vocational Guidance Quarterly,* 22: 124 - 129, 1973.

ACHIEVING HOME-SCHOOL CONTINUITY IN THE SOCIALIZATION OF AN ACADEMIC MOTIVE[1]

ROSEMARY SWANSON
RONALD W. HENDERSON
University of Arizona

ABSTRACT

A field experiment was conducted to test the effectiveness of procedures designed to enable parents to influence children's motivation toward reading, a goal commonly valued by parents and school personnel. Twenty Papago native American mothers were trained to use reinforcement principles to increase the positive valence of reading as a free-choice activity for their second grade children. Analysis revealed a significant increase in children's preference for reading materials in a structured free-choice situation. A generalization trial utilizing a control group was designed to determine if training effects generalized to a classroom free-choice situation. Analysis revealed a significant difference in preference for reading favoring experimental over control group children. Implications for cooperative home-school efforts are discussed.

DURING THE PAST DECADE legislation has spawned a variety of programs intended to deal with the fact that the schools have been less effective in educating children from minority cultural backgrounds than in meeting the needs of the middle-class children whose parents generally design and operate the educational programs in this country. There is little agreement among psychologists and educators even about the appropriate questions to be asked and assumptions to be employed in seeking to remedy the differential effectiveness of educational programs for various sub-cultural groups.

Two points of view have dominated the design of innovative programs intended to remedy inequality in educational effectiveness. The first and most influential view has been that culturally different children fail to profit as much from school instruction as their middle-class peers because of intellectual deficits, which are attributed to inadequate intellectual stimulation in the home environment (5, 6, 7). A less widely held but vocally stated point of view asserts that the cause of culturally different children's failure to profit from instruction is that the programs and procedures of the schools are ethnocentric, and therefore inappropriate and misguided when applied to children from non-middle-class backgrounds. Proponents of this view claim that the "inadequate environment" argument serves as the basis for institutional racism, as manifested in compensatory education programs (2).

The present research was designed to examine the viability of a third and alternative set of assumptions. This third viewpoint hypothesizes that discontinuities between the valued objectives and child training practices of the home and school may inhibit the effectiveness of school programs. This point of view seems to offer greater potential for generating productive solutions than the competing sets of assumptions because it suggests that an initial step toward a solution of the problem of unequal educational effectiveness would be to seek objectives which are mutually desirable to school personnel and parents, and to design mutually acceptable strategies to influence children's progress toward these objectives both at school and at home. In this context, continuity is conceptualized as congruence between the objectives for children's learning that are mutually valued by the home and school, and compatibility between the educational practices of the school and the socialization practices in the home which are articulated to those objectives.

Theoretical Perspective

For reasons to be discussed later, the independent variable selected for study in the present investigation was an academic motive, interest in reading materials. The intervention was designed to bring about changes in children's preference of reading materials when these materials were in competition with attractive alternative choices.

The assumptions about motivation which are most widely accepted in education have been derived, usually informally, from theories of personality, which generally conceptualize motivation as an enduring state or energy system which governs the organism's activities (1). Such theories fail to provide guidance for action programs because they leave it unclear how one would go about influencing individuals to adopt new motives. A central question to be addressed in any effort to alter motives is "How do some people come to feel pride in academic accomplishments, while others feel no sense of satisfaction from such deeds?"

One view (13) suggests that when an event that has already acquired reinforcement value, such as praise, approval, or material incentives, is paired with a class of behaviors such as academic accomplishments, the behaviors associated with striving for academic success may acquire secondary reinforcement value. Since direct reinforcement for academic behaviors is delivered on an intermittent basis in the natural environment, such activities are likely to be maintained for quite some time in the absence of external reinforcement, and the observer who witnesses this behavior may well conclude that the learner is intrinsically motivated.

Support for this interpretation is provided by Winterbottom's descriptive study of achievement motivation in the eight-year-old boys (15). He found that mothers of the boys with high motivation made earlier demands on their sons for independence and excellence than did mothers of boys with low achievement motivation. Moreover, when the boys with high achievement motivation met maternal expectations, their mothers reported using physical affection more frequently than did mothers of the less motivated boys.

Given these well-established theoretical considerations and support from descriptive research, there has been surprisingly little experimentation to validate experimentally the conditions which are thought to influence the development of academic motivation in young children. The behavior modification literature is replete with studies in which children have been influenced to increase the amount of time they spend on academic tasks, or to be more persistent at academic activities. But to the knowledge of the present authors, the on-task behavior in all of these studies has been maintained through the use of externally imposed contingencies, e. g., (13). While incentives of various kinds may play an important role in school learning, it is important to find ways to influence children to pursue learning activities in the absence of external control imposed by teachers or others. An important long-range goal of parents and educators alike is to establish conditions that will result in the acquisition or relatively enduring preferences for activities that may lead to further learning.

Past research has established that there is a strong relationship between home influences and school learning (4, 16, 10) and that this relationship is an impressively stable one (8). Furthermore, research suggests that the degree to which parents demonstrate that they value language and school-related behavior is highly associated with school achievement (9). Since home variables account for a major portion of the variance in children's school achievement, it seems reasonable to attempt to determine if differential parental reinforcement of children's choices of activities will lead to an increase in the frequency with which children select reading materials as a preferred activity, and if this influence will generalize to the classroom.

From the theoretical literature, descriptive research on achievement motivation, and studies of the relationship of academic accomplishment to home environments, the picture which seems to emerge is one in which children who develop motivation toward reading encounter reading activities within a wide range of established relationships with parents, siblings, grandparents, and other significant people. In general, they have dependent relationships with most of these people, resulting in a range of positive affective associations to cues in the setting in which these events take place. Consequently, the child who has had such a background has experienced great redundancy in expressions of value toward reading. Expressions of "value reading" are in the form of models who read, and direct approval to the child for engaging in activities related to reading. This class of events may be thought of as an invariant occurring within a great range of experiences in which other factors (nurturant people, situations, specific nature of the reading materials) are randomly varied. In addition to this heavy redundancy within a highly varied range of events, most of these experiences take place in situations of positive affect. It would come as no surprise, then, that a "motive for reading" acquired in this context would have a high probability for generalization to the school environment.

This conceptualization of the development of motivation toward reading guided the design of the intervention tested in the present research.

Goal Selection

Neither the goal of promoting continuity between home and school influences on the development of an academic choice or motivation in the children involved in this study, nor the specific objective of developing positive motivation toward reading activities, was selected arbitrarily. The present research represents an extension of an earlier parent-training program (11, 14) in which parents were successfully trained in the use of specific socialization practices designed to facilitate the development of question-asking skills in their first grade children. At the conclusion of that program, the Title I Parent Advisory Committee which had initiated the first study requested a continuation of the program, with new objectives focusing on some aspect of children's reading behavior.

The program which was designed in response to this request aimed at teaching parents who were trained during the previous year to generalize their skills to a new set of responses in their children. The specific objective was to train parents to influence their children to be more interested in reading and to evidence this change in motive by choosing reading materials with increasing frequency during free-choice time. Changes in strength of motivation relating to reading activities were to be assessed by examining children's choices of reading materials in comparison with other attractive alternatives in a standardized free-choice situation, and by testing the generalization of the preference to a classroom situation.

It was hypothesized that children whose mothers were trained in procedures designed to influence children's

activity preferences would (*a*) show an increase in their selections of reading materials over attractive alternatives in a standardized free-choice situation, and (*b*) display generalization of this preference to the classroom, by choosing reading activities with significantly greater frequency than a control group of classmates.

Method

Subjects

The participants in this experiment were twenty native American Papago second grade children. All children were enrolled in two elementary schools on the Papago Indian Reservation in Arizona. These subjects were twenty of an original thirty whose mothers had been involved in a training program during the previous year. While the original thirty children had been randomly selected, the twenty involved this year were those whose mothers were able to continue training for a second year.

In order to keep training groups small enough to provide the individualized attention, the mothers of the twenty subjects were randomly assigned to two treatment groups. Training for one group was completed before training for the second group began.

Procedure

Mothers of all twenty *S*s were trained to employ social learning principles to influence their children's choice behavior by increasing the positive valence of reading materials. Two Papago women were trained to serve as parent trainers. Both women were bilingual and hence able to conduct teaching sessions in the primary language of the participating mothers. Through a combination of the use of modeling procedures, role-play, and prepared written lesson plans, the paraprofessionals were taught to model the desired behavior of a mother and child informally examining and discussing reading material. Furthermore, they were trained to demonstrate how a parent might reinforce the child for attending to the reading materials so that his interest could be sustained; and finally the paraprofessionals learned to demonstrate the use of verbal praise that parents might employ with their child when they engaged in self-initiated reading activity during a free-choice situation. The paraprofessional women did not begin training the parent groups until they could perform and demonstrate mastery of these behaviors.

Immediately following the training for the paraprofessional change agents, parent training was initiated. Training was completed for the first ten mothers before training was initiated for the second group. With the first group of mothers, usually six to eight attended group training sessions, while the remainder were trained individually in their homes. Because many mothers in Group II were employed, only two or three were able to attend group sessions regularly, and the remainder had to be trained individually. All group training sessions were monitored by a member of the project research staff, but the paraprofessionals had primary responsibility for the training.

Instruction was divided into five lessons. The first lesson involved teaching the parents the appropriate parent-child interaction sequence which involved a mother and child informally examining and discussing a book together. The paraprofessionals modeled the desired parent-child interaction for a ten-minute period. Following the trainer modeling sequence, the mothers were divided into pairs and would then role-play the interaction sequence. Each mother was given the opportunity to role-play both the mother and child part, and role-play was continued until all mothers were able to perform the desired behaviors. Mothers were then instructed to conduct two training sessions with their child prior to the second training group meeting. At this point each parent-child session involved (*a*) performing the desired interaction sequence with their child for a ten-minute period, and (*b*) observing the children for one hour following the session and recording the amount of time the child engaged in self-initiated activity with the reading materials.

Each subsequent training session increased the complexity and number of behaviors that the mothers were required to master. However, only one new behavior was added each session to avoid confusion and to insure mastery prior to the presentation of an additional novel behavior (see Table 1). Lesson 2 added the use of verbal praise for attending behaviors during the mother-child interaction sequence, and Lesson 3 added the utilization of reinforcement for child-initiated reading activity during the one-hour observation period that followed the ten minutes of mother-child interaction.

Since the goal of the study was to increase children's preference for reading activities over competing attractive alternatives in a free-choice situation, it was necessary to include opportunities for decision- and choice-making during the training. Consequently, in Lesson 4 parents were trained to introduce their children to a novel and attractive additional stimulus and to allow the children to examine the stimulus prior to mother-child interaction with the reading materials. In Lesson 5 an additional novel stimulus was added which provided the child with a total of three alternative activity selections. In both cases, parents were taught to deliver verbal reinforcement to their children for the selection of reading materials over the other during the one-hour observation period.

In this way children were presented with reading materials in a highly valenced situation (i. e., interaction with the mother), but additionally they were provided with the opportunity to make choices and to be reinforced for the selection of reading materials over more novel and visually attractive stimuli.

Data to evaluate the effects of the intervention were taken under two conditions. The first condition was a

Table 1.—Content of Parent Training Session

Lesson	Behavior in Training	Parent-Child Session
1	1. Parent-child role-playing sequence with reading material stimulus	1. Ten minutes of interaction with reading materials. 2. Following parent-child session, observation and recording of child-initiated reading activity during a one-hour period
2	1. Parent-child role-playing sequence 2. Role-play use of verbal praise with child for attention to reading stimulus	1. Ten minutes of interaction with reading materials 2. Use of verbal praise during session for attending behaviors 3. Observation and recording of child-initiated reading activity during a one-hour period
3	1. Parent-child role-playing sequence 2. Role-play use of verbal praise with child for attending behaviors 3. Role-play use of praise during observation period for child-initiated reading activity	1. Ten minutes of interaction with reading materials 2. Use of verbal praise during session for attending behaviors 3. Observation and recording of child-initiated reading activity during a one-hour period 4. Use of verbal praise for reading activity during observation period
4	1. Parent-child role-playing sequence 2. Role-play use of verbal praise with child for attending behaviors 3. Introduction of a second choice stimulus (puzzles) 4. Role-play use of praise during observation period for child-initiated reading activity	1. Ten minutes of interaction with reading materials 2. Use of verbal praise for attending behaviors 3. Introduction of child to puzzle stimulus 4. Observation and recording of child-initiated reading activity with puzzle stimulus present 5. Use of verbal praise for reading-choice activity during observation period
5	1. Parent-child role-playing sequence 2. Role-play use of verbal praise with child for attending behaviors 3. Introduction of a third-choice stimulus (blocks) 4. Role-play use of praise during observation period for child-initiated reading activity	1. Ten minutes of interaction with reading materials 2. Use of verbal praise for attending behaviors 3. Introduction of child to block stimulus 4. Observation and recording of child-initiated reading activity with puzzle and block stimuli present 5. Use of verbal praise for reading-choice activity during observation period

situational task in which children's choice behaviors were measured under stimulus circumstances similar to those which had been used in procedures employed by the child's mother during home intervention. These data were intended to show changes in the choice behaviors of children whose parents were trained, and were collected before and after parent intervention. The second condition was a free-choice time in the child's classroom, which provided a measure of generalization effects to a new situation. A control group was employed in this condition.

For each trial under the situational task condition, each child was taken individually to a small room by a Papago *E*. The room contained three tables, upon each of which was displayed an array of materials which included reading materials (trade books), blocks, and puzzles. Placement of the stimuli on the tables was randomized for each child. The children were invited to play with any of the materials they wished, and they were free to change their choice at any time. Observational data were collected for a ten-minute period during which the *E* sat unobtrusively in a corner of the room. At each ten-second interval the observer marked a check-sheet to note in which of the activities the child was engaged. Scores were expressed in ratio form, and converted to decimal fractions by dividing the number of initiations with reading materials by the total number of activity initiations. The stimuli available for choice were similar to the ones used during training with parents.

For the generalization condition three kinds of activity centers were arranged in the rear of the classrooms in which the sample children were enrolled. Again, each child was individually invited to interact with the stimuli, and behavior was observed with the interval recording observation schedule. A control group was randomly selected from each classroom and similarly invited to interact with the materials.

Table 2.–Descriptive Statistics

	For Situational Task					
	Pre-test			Post-test		
Group	*N*	$\bar{X}$	*SD*	*N*	$\bar{X}$	*SD*
I	10	.19	.18	10	.45	.35
II	10	.18	.18	10	.35	.32

For Generalization Task

Group	*N*	$\bar{X}$	*SD*
Experimental	20	.35	.18
Control	19	.25	.26

Table 3.–Summary of Analysis of Variance

Source	*df*	*MS*	*F*
Between Groups			
Groups	1	.03	<1.00
Error	18	.072	
Within Groups			
Trials	1	.47	5.875*
Trials X Groups	1	.016	<1.00
Error	18	.08	

* $p < .05$

Results

Descriptive data on situational task performance is presented in Table 2.

Situational task data were analyzed using a 2 (groups) $\times$ 2 (trials) repeated measures analysis of variance. Analysis revealed a significant trials effect ($df = 1.18$, $F = 5.88$, $p < .05$) for the experiment. Post hoc analysis conducted on the group effects revealed no significant group differences across testing points. Table 3 presents a summary of the analysis of variance.

The generalization test was evaluated with an independent group *t*-test, and a significant difference favoring the experimental group was revealed ($df = 37$, $t = 5.2$, $p < .01$). Intercoder reliability for the observation instrument was 99% as determined by examination of the number of observations in agreement divided by the total number of observations recorded.

Data from the recording sheets which parents maintained to note the amount of time the children engaged in self-initiated reading activity during the one-hour observation period revealed the following (*1*) 78% of the children engaged in reading activity for thirty minutes or more during the one-hour observation period, by the end of the training period; (2) 71% engaged in self-initiated activity for forty-five minutes or more; and (3) 21% of the children engaged in sixty minutes of self-initiated reading activity.

Discussion

The goal of the parent-training program reported here was to increase the valence of reading materials and hence to increase the frequency of reading material selection in the free-choice situation. Examination of the significant effects for trials would indicate that this was indeed accomplished for children of parents in both training groups. By the end of training, nearly half of these children's self-initiated selections were for reading materials. In addition, the amount of time the children engaged in reading activity

at home during training was impressive. Furthermore, on the classroom generalization task, experimental children selected reading materials more often than control children whose parents had not experienced training.

A question that rises from inspection of the data is why, at the final testing phase, the second group of children did not attain as high a level of performance as did children whose parents were in the first group. While this difference was not statistically significant, it is suggestive. A much higher percentage of mothers of Group I (60-80%) were regular in attending group training sessions. In Group II, however, only 20-30% were in regular attendance, the remainder being trained at home by the paraprofessionals. It may well be that group training sessions are more effective for a program such as this than home visitations. Group training allows for the possibility of peer modeling influence as the mothers have ample opportunity to observe each other as well as the paraprofessionals. Therefore, they are exposed to multiple-model input and accompanying repetition. This would not be possible in the course of training individually conducted in the home. Second, group sessions allow for careful monitoring by a research staff member. While not directly carrying out instruction, the researcher was able to observe parental behavior, note when criterion level was reached, and provide corrective feedback when necessary. Monitoring is not possible during home training.

Implications

The results of this study suggest that the procedures employed provide an effective means for parents to facilitate the development of a motive in their children which is important to both the home and school. The study identified a possible starting point for achieving home-school continuity, but certainly motive-creation of any practical magnitude is a very complex task (3). Academic motivation is developed slowly through experiences at home and school, and where those environments are very disparate, the lack of continuity between home and school experiences may inhibit the development of motivation for academic activities. For this reason, the present research was followed by a feasibility study involving a larger range of child behaviors which could become the focus for cooperation and coordinated effort between teachers and Papago parents. Arrangements were made for each parent and the teacher of their child to define cooperatively a target behavior for the individual child and, with the help of project staff, to devise and implement an intervention plan which could utilize skills learned by the parent through prior participation in the program.

Approximately half of the mothers were involved in this pilot effort, and additional insights have been gained which should serve to guide future attempts to influence home-school continuity. In the "cooperative goal-setting" it seems clear that choices were primarily predicated on teacher goals and priorities, agreed to by parents. Parental knowledge of the school situation was very limited, and it would be unrealistic to expect parents to act assertively in an unfamiliar and somewhat intimidating environment. Teachers were equally unfamiliar with the children's out-of-school behavior and capabilities. It appears that a necessary step for any comprehensive movement toward increased continuity between home and school in cultural settings such as this is for parents and educators to gain greater familiarity with each other's goals. In all likelihood this must be accomplished in individual face-to-face interactions, because while representative bodies such as parent advisory boards do exist, Papagos are very reluctant to presume to represent the opinions of other parents. It is common to hear a representative say, "I don't know how others may feel, but as for me. . .."

Parents may be successfully trained in skills required to influence the intellectual competencies and specific motives of their children, but application of those skills to reach a significant number of jointly endorsed goals under conditions of severe discontinuity requires sustained effort and careful planning.

Summary

Discontinuity in goals and socialization practices of Anglo-dominated schools and the homes of children from ethnic minorities is a problem of national concern. A field experiment was conducted to test the effectiveness of procedures designed to enable parents to influence children's motivation toward reading, a goal which provided a point of continuity in the values of parents and school personnel. Two groups of ten Papago native American mothers each were trained at different points in time by Papago paraprofessionals to use reinforcement principles to increase the positive valence of reading as a free-choice activity for their second grade children. Effects of parent intervention were analyzed with a 2 (groups) × 2 (trials) repeated measures analysis of variance. The main effect for trials was significant ($p < .05$), and post hoc tests revealed no differences between training groups. A generalization trial utilizing a control group was designed to determine if training effects generalized to a classroom free-choice situation. An independent group t-test revealed a significant difference favoring the experimental group children over the controls ($p < .01$).

A follow-up feasibility study explored possible ways of using skills learned by parents in the training program to expand the base of home-school continuity. Implications of descriptive findings were discussed.

FOOTNOTE

1. The work reported herein was conducted under subcontract to Indian Oasis School District #40, Arizona. The project was supported by the Arizona State Department of Education as Project No. 74-912C, under the authority of P. L. 89-10, Title I.

The opinions expressed in this report do not necessarily reflect the position or policy of the Indian Oasis School District, or the Arizona State Department of Education, and no official endorsement by these agencies should be inferred.

Appreciation is expressed to the Papago parents and children who participated in the research, and to the school administrators and teachers who cooperated in the effort. We also wish to gratefully acknowledge the contributions of Irma Dean Edmond, Elizabeth Siquieros, and May Galvez for their contributions to the field work. Thanks are also expressed to Jean Godier, who provided secretarial services to the project and typed the final manuscript.

REFERENCES

1. Bandura, A., *Behavior Modification,* Holt, Rinehart and Winston, New York, 1969.
2. Baratz, S. S.; and Baratz, J. C., "Early Childhood Intervention: The Social Science Base of Institutional Racism," *Harvard Educational Review,* 40:29-50, No. 1, 1970;
3. Brown, R., *Social Psychology,* The Free Press, New York, 1965.
4. Dave, R.H., "The Identification and Measurement of Environmental Process Variables That Are Related to Educational Achievement," unpublished doctoral dissertation, University of Chicago, 1963.
5. Deutsch, M., "Facilitating Development in the Pre-school Child: Social and Psychological Perspectives," *Merrill Palmer Quarterly,* 10:249-264, No. 3, 1964.
6. Hunt, J. McV., *Intelligence and Experience,* Ronald Press, New York, 1961.
7. Hunt, J. McV., "The Psychological Basis for Using Pre-school Enrichment as an Antidote for Cultural Deprivation," *Merrill Palmer Quarterly,* 10:209-248, No. 3, 1964.
8. Henderson, R. W., "Environmental Predictors of Academic Performance of Disadvantaged Mexican-American Children," *Journal of Consulting and Clinical Psychology,* 38:297, No. 2, 1972.
9. Henderson, R. W.; Bergan, J. R.; and Hurt, M., Jr., "Development and Validation of the Henderson Environmental Learning Process Scale," *Journal of Social Psychology,* 88:185-196, 1972.
10. Henderson, R. W.; and Merritt, C. B., "Environmental Backgrounds of Mexican-American Children with Different Potentials for School Success," *Journal of Social Psychology,* 75: 101-106, 1968.
11. Henderson, R. W.; and Swanson, R., "The Application of Social Learning Principles in a Field Setting: An Applied Experiment," *Exceptional Children,* September 1974, 53-55.
12. Staats, A. W., *Learning, Language and Cognition,* Holt, Rinehart and Winston, New York, 1968.
13. Staats, A. W.; Staats, C. K.; Schutz, R. E.; and Wolf, M. M., "The Conditioning of Textual Response Using 'Extrinsic' Reinforcers," *Journal of the Experimental Analysis of Behavior,* 5: 33-40, 1962.
14. Swanson, R.; and Henderson, R. W., "Problems and Issues in Utilizing Research Knowledge in Culturally Diverse Settings," paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1974.
15. Winterbottom, M. R., "The Relation of Need for Achievement to Learning Experiences in Independence and Mastery," in J. W. Atkinson (ed.), *Motives in Fantasy, Action, and Society,* Van Nostrand, Princeton, N. J., 1958.
16. Wolf, R. M., "The Identification and Measurement of Environmental Process Variables Related to Intelligence," unpublished doctoral dissertation, University of Chicago, 1964.

DIFFERENCES BETWEEN HIGH AND LOW ACHIEVERS ON SELF-PERCEPTIONS

BERNADETTE M. GADZELLA
GLENN P. FOURNET
East Texas State University
Commerce, Texas

ABSTRACT

Differences between and changes over a semester on self-perceptions of a quality student were analyzed for 162 high and 120 low achievers. A self-rating scale, devised with 37 student-suggested characteristics of a quality student and designed to record three different ratings on stanine scales, was used to collect the data. Characteristics were collapsed into five general groups reflecting learning in class; study habits and attitudes; peer relationships; student-instructor relationships; and physical and emotional needs. Trend analysis showed significant differences between the two groups on three group characteristics, and significant upward shifts and interaction effects on all five group characteristics. Interesting self-rating patterns emerged. Traits of high and low achievers should be invaluable knowledge for instructors.

IT IS ASSUMED THAT intellectual and scholastic aptitude are prerequisite to success in college. However, some researchers (2, 6) report that measures of academic ability alone are not sufficient factors in predicting the level of academic performance, and other investigators (4) report that there is no significant relationship between

college students' abilities and their levels of academic performances. However, differences in levels of academic performances (whether they are labeled as high and low achievers or over- and underachievers) have been reported to be significantly related to different personality characteristics and behaviors, such as self-perceptions (1, 2), self-confidence, attitudes, motivational drives (2, 5), past performances (1), and strategies used in studying (4, 5).

Alexander (1) stated that the key to a student's success or failure lies in his self-perceptions and how others (e. g., teachers) interpret his performances. When a student perceives himself as a failure, he develops an anxiety which impedes and reflects his performance.

In a study of over- and underachievers, Lum (5) reported that overachievers tend to be more self-confident, have a greater motivation for studies, and greater capacity for working under pressure than underachievers. The underachievers were described as having a greater tendency to procrastinate, to rely upon pressures in completing assignments, and to be more critical of educational methodology and philosophy than the overachievers.

Goldman and Hudson (4) found no significant differences between abilities and high-, middle-, and low-grade-point-average groups, but they did find significant differences among these groups and the strategies the students used in studying. Specifically, these significantly different strategies were found to be *planfulness* (reflecting punctuality and fore-planning, e. g., class meetings, completing notes and assigned readings) and *formal reasoning* (reflecting logical and mathematical reasoning). The investigators supported the idea that these specific study strategies may be more fundamental determinants of college students' levels of academic performances than the students' abilities.

Studies on non-intellectual factors and academic achievement have shed some light on characteristics of different types of students, but additional information is needed in order to indicate which non-intellectual and/or personality trait(s) influence different levels of academic performance.

The present study attempted to examine the problem by employing a different approach from those in the studies reviewed. It investigated (*a*) whether differences exist between high and low achievers on their self-perceptions of a quality student and (*b*) whether any changes occur during a semester for these groups in their self-perceptions of a quality student. (*High* and *low achievers* are referred to here as students who have received grades of A and C, respectively, in a course in which they were enrolled.)

It was hypothesized that there would be (*a*) significant differences between high and low achievers in their self-ratings of a quality student; (*b*) significant upward shifts in the trends over trials during the semester; and (*c*) significant interaction effects between high and low achievers on characteristics of a quality student over trials.

Method

Subjects

The *S*s were 282 students (from different major fields of study) enrolled in introductory, educational, adolescent, applied, and industrial psychology, and clinical and learning theory classes. Several instructors (male and female) taught the classes using their own styles and methods of teaching (e.g., lecture, discussion, group and individual projects and presentations). *S*s participated on a voluntary basis. *S*s were not aware if (or how) the data were going to be analyzed but were assured that their ratings would not in any way influence their course grades. Those *S*s who rated themselves on all the characteristics of a quality student during each of the three rating sessions were included in the study.

After the semester ended, course grades were added to the students' self-rating scales. From this group, students with course grades of A and C (high and low achievers, respectively) were selected. The *S*s were 162 high achievers and 120 low achievers.

Instrument

The instrument used to collect the data was a self-rating scale with 37 student-suggested characteristics of a quality student. The instrument was designed so as to enable students to record three different ratings on stanine scales for each of the characteristics. The scales were numbered 1-9, with "1" indicating the lowest (least desirable) level, "5" the middle (neither desirable nor undesirable), and "9" the highest (most desirable) level.

In the instrument, the characteristics were grouped into two categories: *In Class* and *Out of Class*. Nine of the 37 characteristics, those related to classroom instruction and learning, were placed in the *In Class* category. These characteristics were: attended classes; came prepared to classes; was alert and attentive; participated in class discussions; was open minded; was interested in the subject matter; understood course objectives; took good notes; and asked when I did not understand material.

The remaining characteristics, in the *Out of Class* category, were grouped under four sections: Study Habits and Attitudes; Student-Student Relationships; Student-Instructor Relationships; and Physical and Emotional Needs. Fifteen characteristics fell under the Study Habits and Attitudes section; they were: had a good study schedule; had a special place for study; had a positive attitude toward learning; was determined and studied hard; read textbooks and references assigned; was well organized; used library effectively; did extra work for personal satisfaction; avoided hasty decisions; used dictionary effectively; admitted when I learned materials; evaluated myself often; did my best in all assignments; interrelated course contents; and set goals and objectives. There were three characteristics in the Student-Student

Relationships section, namely, discussed topics and ideas; made friends in classes attended; and formed my own opinions. The section Student-Instructor Relationships consisted of the following four characteristics: got to know the instructor; asked instructor for help; respected instructors; and cooperated with instructors. The remaining six characteristics fell in the Physical and Emotional Needs section. Specifically, these characteristics were: had a well-balanced diet; had sufficient amount of rest; learned to relax; participated in physical activities; did not overload myself with work; and developed other interests and hobbies.

Procedure

Students were informed as to the meaning of the 1–9 numbers on the scales and given instructionas as to how to record their responses. On three different occasions during the semester (beginning, mid-term, and end), students rated themselves on the 37 characteristics.

The administration of the self-rating scales was done during class periods, allowing as much time as the students needed to complete it. The three self-rating sessions were similar with one exception. During the mid-term and final rating sessions a separate self-rating form was used. The use of the separate form was to provide students an opportunity to record their perceptions at that time without being influenced by their previous ratings. Then, the ratings from the special form were transferred to the initial rating scale.

For purposes of analysis, the ratings for each of the previously mentioned characteristics in the *In Class* and the sections in the *Out of Class* categories were collapsed (Table 1). A trend analysis (3) was used to analyze the data in which the high and low achievers (groups) were treated as the main effects and the three ratings on each group of characteristics as the trial effects. Post-*t*-tests were computed for the trial and interaction (main × trial) effects for the groups of characteristics yielding significant *F*-ratios. The analysis of the post-*t*-tests for the interactions consisted of first adjusting the cell means by column and row effects.

Results and Discussion

The means and standard deviations for high and low achievers for each of the grouped characteristics of a quality student on the three trials are presented in Table 1. In Table 2, the *F*-ratios and their levels of significance are reported for high and low achievers under the column *Group:* trend effects under the column *Trial;* and interaction effects under the column *Interaction.* All post-*t*-

Table 1.–Means and Standard Deviations for Grouped Characteristics of a Quality Student on Three Trials for 162 High and 120 Low Achievers

CHARACTERISTICS OF ACHIEVERS		TRIALS					
		First		Second		Third	
		$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*
In Class	High	7.071	1.000	7.501	0.821	7.842	0.847
	Low	6.697	1.092	6.952	1.012	7.164	1.153
Out of Class							
Study Habits and Attitudes	High	6.125	1.302	6.716	1.102	7.224	0.986
	Low	5.860	1.485	6.114	1.381	6.559	1.461
Student-Student Relationships	High	6.416	1.605	7.068	1.308	7.648	1.104
	Low	6.242	1.777	6.600	1.708	6.881	1.709
Student-Instructor Relationships	High	6.770	1.377	7.321	1.117	7.749	1.030
	Low	6.856	1.444	7.071	1.179	7.233	1.370
Physical and Emotional Needs	High	6.245	1.510	6.591	1.489	7.017	1.334
	Low	6.329	1.531	6.513	1.595	6.529	1.598

Table 2.—*F*-Ratios for Groups (High and Low Achievers), Trials, and Interactions (Groups and Trials) on Ratings of the Grouped Characteristics of a Quality Student

GROUPED CHARACTERISTICS	*F*-RATIO		
	Group *df* 1/280	Trial *df* 2/560	Interaction *df* 2/560
In Class	29.97*	63.66*	3.49**
Out of Class			
Study Habits and Attitudes	14.64*	101.34*	5.25*
Student-Student Relationships	9.49*	62.27*	5.57*
Student-Instructor Relationships	3.37	47.93*	8.10*
Physical and Emotional Needs	1.05	23.66*	7.17*

*$p < .01$ **$p < .05$

tests reported as significantly different in the discussion of the report were significant at .05 level or lower.

Differences between High and Low Achievers

Three of the five groups of characteristics were significantly different between high and low achievers. To this degree, the present hypothesis was confirmed. The three groups of characteristics, revealing significant differences between the groups, were the *In Class* category and two sections of the *Out of Class* category (Study Habits and Attitudes, and Student-Student Relationships). In each case, the high achievers rated themselves significantly *higher* than the low achievers.

Differences over Trials

The results of the analysis (Table 2) show that all of the groups of characteristics yielded significant differences ($p < .01$) over the trials. This was interpreted as both groups having indicated that they had changed over the semester on their self-perceptions of a quality student. It is obviously difficult to determine which factor(s) influenced the changes, whether they were due to students' better conceptions of what characterizes a quality student, or due to changes that occurred because of the influence of various factors, such as familiarity of the environment, better acquaintance with peers and instructors, and/or better knowledge of the subject matter. The fact that changes reflecting upward trends occurred confirms the hypothesis made.

Differences in the Interaction Effects

All five groups of characteristics produced significant *F*-ratios in the interaction effects (Table 2), thus confirming the hypothesis made. In order to perceive more clearly the patterns which emerged between the high and low achievers on the five groups of characteristics, adjusted mean scores for the three trials on each of the five groups of characteristics are graphically displayed in Figure 1.

From these data (Figure 1) one perceives that, generally, high achievers rated themselves low on the initial rating on these characteristics and then higher on the second and

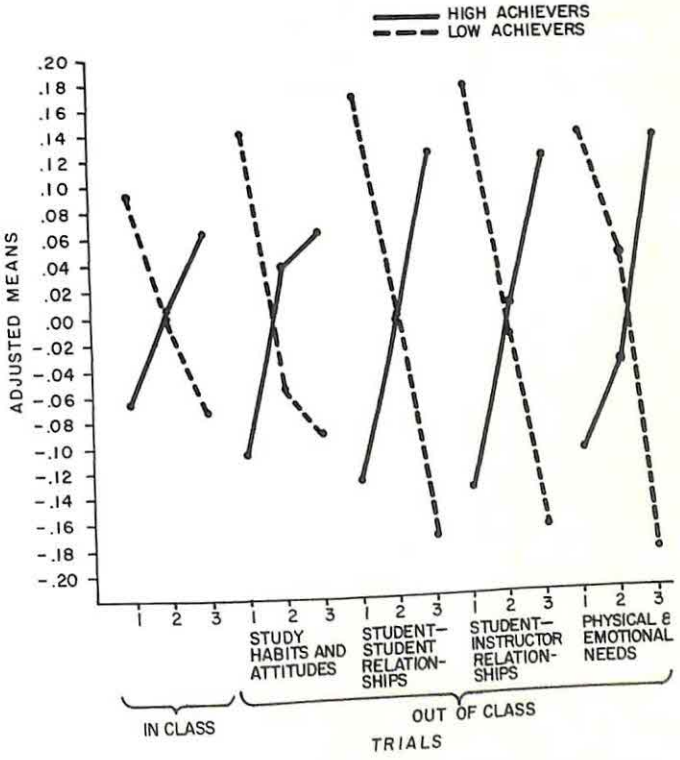


Figure 1.—Adjusted Means on Three Trials for Characteristics of a Quality Student

third ratings, respectively. For the low achievers, the self-rating pattern was just the reverse. Low achievers rated themselves high on the first rating and significantly lower on the second and third ratings, respectively.

Examination of the specific traits in each of the grouped characteristics seems to indicate that attitudes toward learning may be the overall factor influencing a student's level of academic performance. Attitudes are closely related to such characteristics as attending classes, coming prepared for classes, alertness and attentiveness, etc. (listed in the *In Class* category). Attitudes towards learning are also influenced by one's completion of reading assignments, self-confidence, best efforts, objectives and goals, etc. (listed in the Study Habits and Attitudes section). The fact that these traits characterized the high and low achievers in a reverse manner during the semester suggests that instructors might take note of these traits in an effort to help students who are potential low achievers. Other approaches that instructors might implement are to help all students develop positive attitudes towards learning, assist them in setting specific objectives and long-range goals, and assist them in making frequent self-evaluations.

Previous studies (1) have shown that peer and student-instructor relationships are related to academic performances. The present study reveals additional information in that the interactions over the semester for peer and student-instructor relationships operate in a reverse fashion for high and low achievers.

Undertaking reasonable study loads, learning to relax, and participating in physical activities are factors which reduce anxiety and tension. Counseling students on the realities should reduce, or even eliminate, feelings of inadequacy and tension and aid in developing a more positive attitude toward learning and academic performance.

Conclusion

The findings in the present study are in agreement with those of previous reports (1, 2) that students' self-perceptions are related to their level of academic performance. In addition, the present study shows, more specifically, how self-perceptions of a quality student by high and low achievers differ and change over the semester. Knowledge of the traits descriptive of high and low achievers should hlep instructors in better understanding their students. As a result, instructors might take a positive approach by teaching students how to study, help students develop positive attitudes towards learning, assist students in setting specific objectives and long-range goals, and assist students to make frequent self-evaluations. Being cognizant of the fact that some students will need more help than others, an instructor may then provide experiences in which students can succeed and thus assist these students in developing a better perception of themselves. These approaches may be a beginning in a direction leading to higher academic performances.

REFERENCES

1. Alexander, E. D., "Marks and Their Effects on Poor Achievers," *Teachers College Journal*, 36:110-113, 1964.
2. Bowman, P. H., "Personality and Scholastic Underachievement," in D. E. Hamachek (ed.), *Human Dynamics in Psychology and Education, Selected Readings*, Allyn and Bacon, Boston, 1972, pp. 66-78.
3. Edwards, A. L., *Experimental Design in Psychological Research*, (Fourth Ed.), Holt, Rinehart and Winston, New York, 1972.
4. Goldman, R. D.; and Hudson, D. J., "A Multivariate Analysis of Academic Abilities and Strategies for Successful and Unsuccessful College Students in Different Major Fields," *Journal of Educational Psychology*, 65:364-370, 1973.
5. Lum, M. K. M., "A Comparison of Under- and Overachieving Female College Students," *Journal of Educational Psychology*, 51:109-114, 1960.
6. Smart, J. C.; Elton, C. F.; and Burnett, C. A., "Underachievers and Overachievers in Intermediate French," *Modern Language Journal*, 54:415-420, 1970.

ANALYSIS OF THE UNIT TESTING COMPONENT OF THE PERSONALIZED SYSTEM OF INSTRUCTION

JANICE MACLIN
ROBERT WILLIAMS
University of Tennessee

LINDA CLARK
Wake Forest University

ABSTRACT

The present study compared examination scores of 173 undergraduate students in a course taught by the Personalized System of Instruction under three conditions: required unit testing, optional unit testing, and no unit testing. Each of the six course sections was randomly assigned a different testing sequence in order to determine whether the success of PSI had been due to unit testing procedures or to the study questions. The dependent variable was a 65-item multiple choice examination administered at the end of each three-week phase of unit testing. The results of a Lindquist Type I ANOVA indicated a significant sequence by conditions interaction ($p < .05$). Tests of simple main effects showed that when required unit testing came first in the test sequence, the scores on the subsequent exam were significantly higher than the scores on the same exam following the other two unit test conditions. A Lindquist Type III ANOVA revealed a significant interaction between GPA and unit test conditions. Tests of simple main effects indicated no significant difference among the three unit test conditions on exams of students with GPAs in the upper 25 % of the class. Students with GPAs in the lowest 25%, however, attained their highest scores on the exam which followed the required unit testing, and scored higher on the exam following no unit testing than on the exam following optional testing.

AS THE KELLER PLAN (7), referred to as the Personalized System of Instruction (PSI), gains increasing attention, its merits are being evaluated. The PSI method is based on identifying the key elements of a course with study questions from which unit tests and course examinations are composed. Each student is required to achieve a predetermined criterion on a series of unit tests before proceeding to the course examinations. Immediately after each unit test, the proctor is available to provide feedback to the student. Evaluative research on PSI has revealed definite characteristics of this system as compared to the lecture method.

Several researchers (2, 9, 13) have reported that students score higher on final examinations in PSI format courses than students in the lecture format courses. It has also been demonstrated that students score higher on essay tests (2, 13) and express more favorable attitudes towards the course with PSI than with the lecture method (9, 11, 13, 14). Furthermore, with personalized instruction there is usually a skewed distribution with many A's and few low grades (1, 3, 4, 7). Recent research indicates that two major components of PSI which increase students' performance are the study questions and unit tests prior to examinations (5, 6, 10, 12).

The effects of unit testing, however, have not been isolated in any of the preceding studies. In order to determine whether the success of PSI has been due to the unit testing procedure or to the study questions, the present study compared student achievement on course examinations after required unit testing, optional unit testing, and absence of testing.

Method

Subjects, Materials, and Course Format

One hundred and seventy-three undergraduates enrolled in six sections of adolescent psychology served as *S*s during the Fall 1973 and Winter 1974 quarters. All sections were taught by graduate teaching assistants. All students were given a course syllabus which organized the reading assignments, audio-visuals, guest speakers, and study questions into nine units. Approximately one week was devoted to each unit.

Each student was then instructed to write out his own grade contract, which could be altered at any time throughout the quarter. The required components of all contracts were four course examinations and three unit tests. Additional credit could be earned through the following activities: book reviews, field projects, class presentations, class attendance, verbal participation in class discussions, and student aid programs. At the end of the quarter, all points were converted to letter grades according to the following scheme: A = 285 points with a minimum of 27 on the three combined unit test scores and a minimum of 210 on the four course examinations, B = 260 points with a minimum of 24 on the three unit test scores and a minimum of 180 on the four examinations; C = 210 points with a minimum of 18 on the three unit tests and 140 on the four examinations. Credit from the optional activities could be added to these test minimums to attain a desired grade.

Table 1.–Linquist Type I Analysis of Variance

SOURCE	*df*	*SS*	*MS*	*F*
Between subjects	172	24849.688	144.475	
Sequence (*A*)	5	2392.125	478.425	3.5577*
Error (*B*)	167	22457.563	134.476	
Within subjects	346	6594.937	19.061	
Tests (*B*)	2	722.375	361.188	22.55*
A × *B*	10	522.832	52.283	3.26*
Error (*W*)	334	5349.730	16.017	
Total	518	31444.625	60.704	

*$p < .01$

Table 2.–Mean Examination Scores after Each Testing Condition

Unit Testing Sequence*	*N*	After Required Testing		After Optional Testing		After No Unit Testing	
		Mean	*SD*	Mean	*SD*	Mean	*SD*
RNO	27	49.44	6.42	45.15	7.36	45.56	7.10
ORN	23	44.87	7.85	41.44	9.31	39.96	7.14
ONR	31	40.58	6.86	41.23	7.66	41.52	8.51
NOR	28	43.68	7.21	41.50	7.93	42.04	8.46
RNO	35	48.71	5.59	43.54	7.07	46.29	6.76
NRO	29	45.79	8.87	43.35	8.87	45.17	7.32

*R = Required Unit Testing

O = Optional Unit Testing

N = No Unit Testing

Unit Tests and Course Examinations

The course was divided into three phases of three weeks each. During one phase the unit tests were required, during another the unit tests were not available, and during the third phase the quizzes were available but the scores were not recorded. The three phases were administered to each of the six sections in a different, randomly assigned sequence.

During the three-week required testing phase, students were given a list of times and places when proctors would be available. The students were then instructed that each of the three quizzes was composed of ten short answer questions randomly selected from a pool of thirty questions included in the syllabus for each unit. Students could take alternate forms of each quiz up to three times to meet the minimum criterion for their contracted grade. Quizzes were administered by graduate student proctors in a designated classroom. Immediate feedback was available, although students were not required to stay for the feedback. The three required quizzes had to be taken before the examination over those three units was given in class. Class time was not utilized to answer or lecture on the study questions.

Students were also informed that the questions on the four major course examinations would come primarily from study questions in the syllabus. Each of the three examinations was composed of 65 multiple choice items, covering three units of materials. The fourth multiple choice exam was a comprehensive test covering most of the course content.

Experimental Design

A Lindquist Type I ANOVA was utilized (8) with the unit testing condition as the repeated measures factor and the six sequences of testing as the between-subjects factor. There were three levels of testing and six levels of sequence group. The data were further analyzed by use of a Lindquist Type III ANOVA with cumulative grade point average as a blocking variable. The high and low 25% of each group were the two levels of GPA.

Results

The results of the Lindquist Type I ANOVA indicated a significant sequence by test condition interaction as shown in Table 1. The three conditions of unit testing were differentially effective depending on the order of the required testing sequence (Figure 1). Due to the presence of a significant interaction, tests of simple main effects were applied. Means, number, and standard deviations for each group are shown in Table 2. Tests of simple main effects and differences between all pairs of means revealed differential effects on student examination scores for the following sequences: RON, ORN, and RNO (R = Required testing, N = No testing, O = Optional testing). For the RON and ORN sequences, the scores on exams which followed the required unit testing were significantly higher than the scores following the other two conditions ($p < .05$). For the RNO sequence, the exam scores following required testing were significantly higher than the scores after the optional testing ($p < .05$). No differences in subsequent examination scores followed the optional and no unit testing conditions. This can be attributed to students not taking tests under the optional testing condition. (Only 2 of the 173 students chose to take unit tests.)

The results of a Lindquist Type III ANOVA indicated a significant interaction between GPA and testing condition as shown in Table 3. Testing conditions were differentially effective for students with high and low GPAs (Figure 2). Means, number, and standard deviations for each group are shown in Table 4. Tests for simple main effects for each level of GPA indicated no significant differences among the three unit testing conditions for students with a high GPA. However, a Newman-Keuls analysis indicated that students within the low GPA group scored significantly higher on the exam taken after required unit testing than on the exams following the other two conditions ($p < .05$). In addition, the exam scores following no unit testing were significantly higher than the scores following optional unit testing ($p < .05$). Since so few students chose to take unit tests under the optional testing condition, it had been expected that the results of the optional and no unit test conditions would be essentially the same. Perhaps the students with lower GPAs

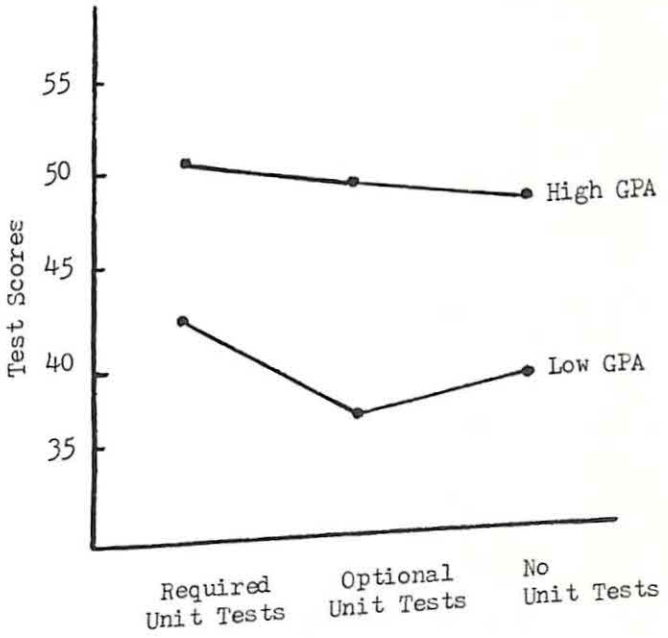
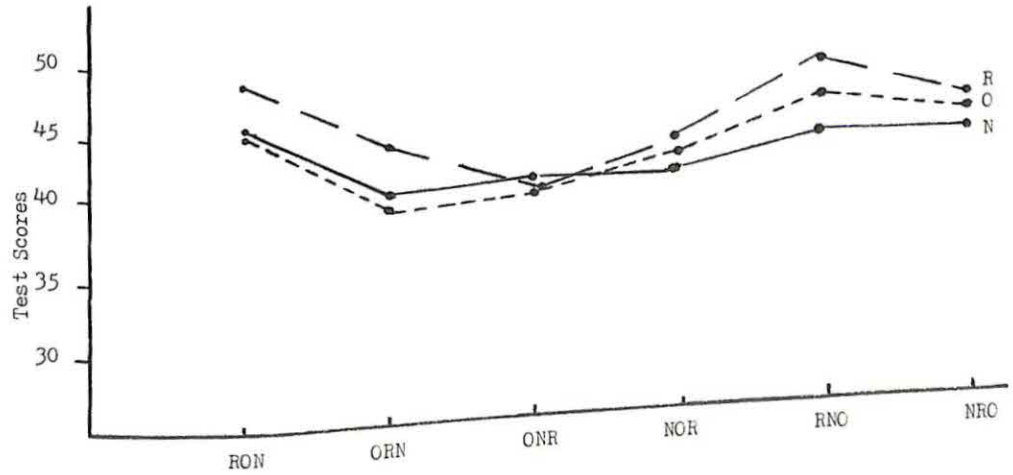


Figure 2.–GPA by Tests Interaction



*R = Required Unit Testing
O = Optional Unit Testing
N = No Unit Testing

Figure 1.–Sequence by Tests Interaction*

Table 3.–Linquist Type III Analysis of Variance

SOURCE	*df*	*SS*	*MS*	*F*
Between subjects	71	12476.398	177.132	
GPA (*B*)	1	5995.563	5995.563	71.74*
Sequence (*C*)	5	1340.004	268.001	3.2067*
B X *C*	5	226.246	45.249	.54
Error (*B*)	60	5014.586	83.576	
Within subjects	144	2964.102	20.584	
Tests (*A*)	2	531.563	265.781	17.00*
A X *B*	2	160.539	80.270	5.135**
A X *B* X *C*	10	144.430	11.443	0.73
Error (*W*)	120	1875.821	15.632	
Total	215	15540.500	72.281	

*$p < .01$ **$p < .05$

Table 4.–Mean Examination Scores After Each Testing Condition for High and Low GPAs

Sequence	GPA	*N*	After Required Unit Testing		After Optional Unit Testing		After No Testing	
			Mean	*SD*	Mean	*SD*	Mean	*SD*
RON	High	6	55.83	3.31	50.00	4.00	51.33	2.58
	Low	6	48.83	5.23	41.33	8.91	42.83	3.43
ORN	High	6	49.00	3.57	47.83	5.57	42.50	9.26
	Low	6	38.00	9.21	32.67	7.84	33.50	4.23
ONR	High	6	48.17	5.98	47.00	7.13	48.50	7.97
	Low	6	38.17	4.07	46.33	5.98	37.67	3.32
NOR	High	6	50.17	6.18	48.33	4.68	48.67	9.54
	Low	6	41.50	5.24	35.00	6.57	35.00	6.89
RNO	High	6	51.00	5.52	49.17	3.06	48.33	3.56
	Low	6	45.83	6.68	37.83	8.84	41.68	8.94
NRO	High	6	49.83	3.25	52.50	4.42	53.50	2.26
	Low	6	41.00	6.23	35.67	7.61	39.17	5.98

*R = Required Unit Testing
O = Optional Unit Testing
N = No Unit Testing

felt more pressure to work under the no testing condition, which resulted in higher exam scores.

These results are in agreement with those of previous researchers (10) who compared exam scores of students who changed from lecture to personalized instruction. In the study by Born et al., students whose scores were in the top half of the class on the first exam were not affected by the change in teaching procedures. The change to personalized instruction, however, produced an increase in exam scores of students in the lower half of the class.

Discussion

There are several implications of the present study for classroom teachers. The most effective method for increasing students' examination scores under a PSI format is to administer a required quiz over the study questions early in the course. Furthermore, students with low GPAs benefit more from required testing than students with high GPAs. Given the choice of whether or not to be given a non-graded quiz over the study questions, the majority of students chose not to be tested, and their subsequent course examination scores did not increase above scores obtained during the required testing phase. The decision not to take unit tests did not markedly decrease the scores of students with high GPAs but did significantly decrease the scores of students with low GPAs. Providing students with study questions did not significantly improve examination scores unless a required unit quiz was administered over the questions prior to the course examination. Teachers, therefore, might explain this phenomenon to their classes in order to prevent students with lower GPAs from making the decision not to be tested over study questions.

Replication of the present study with graduate student or public school populations would further clarify the efficacy of unit tests and study questios. Comparisons might also be made between classes who remain on a required testing schedule throughout a course and classes who are tested only on a final examination.

REFERENCES

1. Born, D.G., *Instructor Manual for Development of a Personalized Instruction Course*, Center to Improve Learning and Instruction, University of Utah, Salt Lake City, 1971.
2. Born, D.G.; Gledhill, S.M.; and Davis, M.L., "Examination of Performance in Lecture, Discussion, and Personalized Instruction Courses," *Journal of Applied Behavior Analysis*, 5: 33-43, 1972.
3. Born, D.G.; and Herbert, E.W., "A Further Study of Personalized Instruction for Students in Large University Classes," *Journal of Experimental Education*, 40: 6-11, 1970.
4. Ferster, C.B., "Individualized Instruction in a Large Introductory Psychology Course," *The Psychological Record*, 18: 521-532, 1968.
5. Jenkins, J.R.; and Deno, S.L., "Influence of Knowledge and Type Objectives on Subject-Matter Learning," *Journal of Educational Psychology*, 62: 67-70, 1971.
6. Jenkins, J.R.; and Neisworth, J.T., "The Facilitative Influence of Instructional Objectives," *Journal of Educational Research*, 66: 254-56, 1973.
7. Keller, F.S., "Good-bye Teacher . . ." *Journal of Applied Behavior Analysis*, 1: 79-89, 1968.
8. Lindquist, E.F., *Design and Analysis of Experiments in Psychology and Education*, Houghton Mifflin, Boston, 1953.
9. McMichael, J.S.; and Corey, J.R., "Contingency Management in an Introductory Psychology Course Produces Better Learning," *Journal of Applied Behavior Analysis*, 2: 79-83, 1969.
10. Miles, D.T.; Kibler, R.J.; and Pettigrew, L.E., "The Effects of Study Questions on College Students' Test Performance," *Psychology in the Schools*, 4: 25-26, 1967.
11. Morris, C.J.; and Kimbrell, G.M., "Performance and Attitudinal Effects of the Keller Method in an Introductory Psychology Course," *The Psychological Record*, 22: 523-530, 1972.
12. Semb, G.; Hopkins, B.C.; and Hursh, D.E., "The Effects of Study Questions and Grades on Student Test Performance in a College Course," *Journal of Applied Behavior Analysis*, 6: 631-642, 1973.
13. Shepphard, W.C.; and MacDermott, H.G., "Design and Evaluation of a Programmed Course in Introductory Psychology," *Journal of Applied Behavior Analysis*, 3: 5-11, 1970.
14. Witters, D.R.; and Kent, G.W., "Teaching without Lecturing: Evidence in the Case for Individualized Instruction," *The Psychological Record*, 22: 169-176, 1972.

THE EFFECT OF DIFFERING CRITERIA FOR UNIT EXAM MASTERY ON COLLEGE TEST PERFORMANCE[1]

EDWIN CARTER
State University of New York at Geneseo

KATHLEEN TELAAK-CARTER
Greece Central School District
Rochester, New York

EUGENE COUTURE
University of West Virginia, Morgantown

PAMELA WRIGHT
State University of New York at Binghamton

ABSTRACT

The influence of differing criteria for receiving a grade of A on unit exams (representing 75% of the course grade) in a college course was investigated. Group 1 students received an A if they scored 90% on each of the ten unit exams. Group 2 students were required to score 100% on nine of the exams. Group 3 students were required to accumulate 90% of the cumulative total points on the unit exams considered collectively. Group 2 students were significantly superior to those of the other groups in: (*a*) their performance on a common final exam (representing 25% of the course grade), and (*b*) their final course grade. Also, there were significantly more retakes among Group 2 students.

THE APPLICATION OF operant conditioning techniques to classroom instruction has been a topic of much recent research [see (8)], particularly in the area of individualized instruction (4).

In most personalized instruction methods, the course material is divided into discrete, but interdependent, units. Students proceed through the units sequentially and are required to demonstrate mastery of the units by taking exams on the unit material. With an emphasis on what pupils know and not on what they don't know, students may retake exams on units in which exam performance indicates insufficient mastery of the material. Most often students are required to pass units with a perfect score, the final grade being determined by the number of units passed. However, sometimes students are not required to pass individual units, their grade being determined instead by the total number of points accumulated on all the unit exams considered collectively. Further, in many instances a final comprehensive exam grade and/or a laboratory grade is averaged in with the grade from the unit exams to determine the final course grade.

Studies have compared personalized courses to traditional courses, e.g., (5) and (1), and have also examined factors which may be contributory to the success of personalized methods. For example, Farmer et al. (2) found that students of proctors performed better than those of non-proctors in personalized courses. Semb et al. (7) found that students did better on unit exams when test questions were similar to study question items. In a study by Johnston and O'Neill (3) unit exam performance levels were found to vary directly with the performance criteria necessary to receive a grade of A. Semb (6) investigated the same variables as Johnston and O'Neill using a different design and response measure. He concluded that a high mastery criterion produced better test performance than a lower criterion.

Studies investigating variables influencing test performance have typically employed within-subject designs and have dealt with undergraduate populations. The present study investigated the role that differing mastery criteria would have on test performance of different groups of students enrolled in the same course. In addition, a graduate course that met only once a week was selected for study in order to extend the empirical findings of personalized courses to include a more "advanced" population meeting on once-a-week schedule.

Method

Subjects

The *S*s were 102 students assigned randomly to three sections of a course in educational psychology at SUNY-

Geneseo. Initially the number of students in Sections 1, 2, and 3 was 36, 34, and 39, respectively. There were two dropouts each for Sections 1 and 2, and three from Section 3. The senior author was the instructor in all three sections.

Procedure

The course reading material was divided into ten weekly units; exhaustive study questions covering the reading were devised by the instructor. Additionally, a large pool of 76 or more test items was composed for each unit. Test items never duplicated study questions but covered the same content area. All study materials were the same for each section. Each section met once a week for three hours. In each class during the initial 45 minutes, students took a written exam on the week's reading material (i. e., one unit). The rest of the time was devoted to general discussion of reading material and pertinent issues. There were no formal lectures.

Units were covered one week at a time and students could not take a unit exam before it was scheduled in class; however, they were permitted to take other exams covering the same material as often as desired by making an appointment with the instructor who administered the retakes in his office. No retakes could be taken after the data on which the next exam was scheduled (i.e., seven days after original), and only one retake could be taken per day. Test and retake items were randomly selected from the pool of test items for each unit, with the provision that no retake item could have appeared on the original exam or on any previous retake that any particular individual had taken. Students were given grades immediately after taking an exam or retakes. However, students could not keep an exam or retake until after the next unit exam had been scheduled. This latter contingency, the large test item pool, and the seven-day limited hold discouraged students from passing potential retake questions on to someone else.

Each unit exam and retake was worth 20 points and was composed of five multiple choice items and five short answer items each worth 2 points. Multiple choice items were graded either 2 or 0, while short answer items were graded 2, 1, or 0. Only the best score for each unit contributed to the final grade.

One week after the tenth unit exam was scheduled, all students took a comprehensive final exam. The final exam was worth 50 points, and no retakes were permitted. It was composed of 25 multiple choice and short answer questions graded just like items on the unit exams. Included with the final, each student received the standard SUNY-Geneseo course evaluation form, in which students could rate the overall quality of the course on 5-point rating scales (1 = unsatisfactory, 5 = excellent).

In the grading of all tests in the course, neither the names of students nor the sections that tests came from were known to the instructor (who did all the grading) until after all exams were graded. Thus, all grading was blind to insure against score biasing.

The course grade represented a 25% weighting of the final exam and a 75% weighting of the unit exams. On the final exam, letter grade equivalents of rounded point totals were ascertained as follows: A = 45 - 50 (90%); B = 44 - 40 (80%); C = 38 - 40 (75%); D = 35 - 37 (70%). As follows, groups differed according to how letter grade equivalents of unit exams were assigned:

Group 1: In order for the 34 students of Group 1 to receive a grade of A, they had to get 90% of the total possible points on a unit exam on *all ten* unit exams (i. e.,

Table 1.–Group Mean Point Total on Unit Exams, Unit Exam Grades, Final Exam Score, Number of Retakes, and Course Grade

Group	*N*	Mean Total Points on Unit Exams	Mean Grade Point Average for Unit Exams	Mean Final Exam Score	Mean Total Retakes of Unit Exams	Mean Course Grade Point Average
1	34	183.0	3.40	46.1	3.9	3.45
	(SD)	(9.4)	(.19)	(1.9)	(2.6)	(.30)
2	32	187.3	3.85*	49.4*	12.4	3.90*
	(SD)	(9.3)	(.20)	(2.1)	(2.8)	(.35)
3	36	185.2	3.45	45.4	2.6	3.40
	(SD)	(8.4)	(.15)	(1.6)	(2.3)	(.22)

*Differed significantly from Groups 1 and 3 ($p < .01$)

18 of 20 maximum points). Grades of B, C, and D required getting at least 18 points on 9, 8, and 7 of the ten unit exams, respectively.

Group 2: The 32 students of Group 2 received a grade of A if they correctly answered all questions (i. e., 100%) on *nine* of the ten unit exams. Grades of B, C, and D required perfect scores on 8, 7, and 6 unit exams, respectively.

Group 3: The 36 students of Group 3 were required to accumulate 90% of the cumulative total points of *all ten* unit exams considered *collectively* (i. e., 180 out of 200 points) in order to get an A. Grades of B, C, and D necessitated the accumulation of 160, 150, and 140 points, respectively (i. e., 80%, 75%, and 70% of the 200 total points on all unit exams).

Results

The groups were compared with respect to the following six variables: (*a*) total points on the weekly exams; (*b*) grades on unit exam portion; (*c*) final exam score; (*d*) number of retakes taken during the course; (*e*) final course grade; and (*f*) course rating. Table 1 contains the mean group values for each of the above variables. Analysis of the data of Table 1 was as follows:

Total points on weekly exams: Only for Group 3 was there a contingency between total points accumulated on unit exams and course grade. Nonetheless, an analysis of variance indicated that there was no difference among the groups in the total number of points earned on the unit exams.

Unit exam grades: Differences in performance on unit exams were assessed by taking each student's grade for that portion of the course and assigning it a numerical equivalent (GPA) for which A = 4, B = 3, C = 2, D = 1, and E = 0. An analysis of variance was then performed, and the groups differed significantly from each other ($F = 61.50$, $df = 2, 99$; $p < .01$). Post hoc *t*-tests indicated that Group 2 had significantly higher grades ($p < .01$) than the other groups, which did not differ significantly from each other.

Final exam score: The groups differed significantly in their performance on the final exam ($F = 42.47$, $df = 2, 99$; $p < .01$). Post hoc *t*-tests revealed that Group 2 performed better than both Groups 1 and 3 ($p < .01$ for each comparison), which did not differ from each other.

Course grade: Differences in overall course performance were evaluated numerically in a manner identical to that used in evaluating unit exam grades. The groups differed significantly from each other ($F = 25.42$, $df = 2, 99$; $p < .01$). Post hoc *t*-tests indicated that Group 2 had significantly higher grades than either of the other groups ($p < .01$ for each comparison), which did not differ from each other.

Table 2 contains a summary of the data analyses performed on the students' course performance.

Table 2.–Summary of Analyses Performed on the Data of Table 1

SOURCE	*df*	*MS*	*F*
Total unit points	2	138.56	1.69
Error	99	81.80	
Unit GPA	2	2.0049	61.50*
Error	99	.1326	
Final exam	2	151.71	42.47*
Error	99	3.57	
Retakes	2	939.70	139.40*
Error	99	6.74	
Course GPA	2	2.25	25.42*
Error	99	.09	

*$p < .01$

Course evaluations: The mean overall ratings given to the course were 4.4 ($SD = .38$); 4.4 ($SD = .29$); and 4.5 ($SD = .19$) for Groups 1, 2, and 3, respectively. The group ratings did not differ significantly from each other.

For all students, a correlation was calculated between the total number of points on unit exams and the final exam score. The obtained correlation of .84 was significantly different from zero ($p < .01$), indicating that students who did well on unit exams also tended to do well on the final exam.

Discussion

The data indicate that the students in Group 2 performed significantly better than students in the other groups as measured by mean unit exam GPAs, final exam score, and overall course GPAs, and in so doing support Johnston and O'Neill (3) and Semb (6) using different populations and time schedules. Thus, students who were expected to perform at higher levels were generally able to do so. Further study is required to determine if the differences were due to greater studying or some other factor.

Interestingly, students in Group 2 did not accumulate significantly more points during the course. Examination of the data indicates that this resulted because students who had already passed the required number of unit exams necessary for a grade of A did very poorly, that is, only got a few points, on one of the ten unit exams. (Most did poorly on exam 10, although some did poorly on one of the other unit exams and passed exam 10.) Students who received A in the other groups consistently tended to miss one or two points on each unit exam and never did disastrously on any one exam. The superior performance of Group 2 students was sufficient to produce significantly better final course grades.

This study supported the findings of Semb (6) that higher criteria produce not only better examination

performances but more retakes as well, indicating that, in general, it may be difficult to avoid failures on unit exams. In this course there was a small but steady decrease in retakes as the course progressed. Perhaps greater care should be given to shaping test performance so as to produce learning without errors and, in so doing, possibly avoiding any bad side effects associated with failure. Further, retakes may entail extra time spent by instructors in administration of courses. In the present study, retakes took about 45 minutes on the average to administer and about 15 minutes to compose. Thus, Group 2 occupied more of the instructor's time than the other groups. In the absence of a technology that can prevent unit exam failures, instructors may wish to weigh the advantages of the method used for Group 2 against the extra time required for its administration.

FOOTNOTE

1. Requests for reprints may be sent to Edwin Carter, Department of Psychology, SUNY- Geneseo, Geneseo, N. Y. 14454.

REFERENCES

1. Born, D. G.; Gledhill, S. M.; and Davis M. L., "Examination Performance in Lecture-Discussion and Personalized Instruction Courses," *Journal of Applied Behavior Analysis,* 5: 33-43, 1972.
2. Farmer, J.; Lachter, G. D.; Blaustein, J. J.; and Cole, B. K., "The Role of Proctoring in Personalized Instruction," *Journal of Applied Behavior Analysis,* 5: 401-404, 1972.
3. Johnston, J. M.; and O'Neill, G., "The Analysis of Performance Criteria Defining Course Grades as a Determinant of College Student Performance," *Journal of Applied Behavior Analysis,* 2: 261-268, 1973.
4. Keller, F. S., "Goodbye, Teacher. . .," *Journal of Applied Behavior Analysis,* 1: 79-89, 1968.
5. McMichael, J. S.; and Corey, J. R., "Contingency Management in an Introductory Psychology Course Produces Better Learning," *Journal of Applied Behavior Analysis,* 2: 79-83, 1969.
6. Semb, G., "The Effects of Mastery Criteria and Assignment Length on College-Student Test Performance," *Journal of Applied Behavior Analysis,* 7: 61-69, 1974.
7. Semb, G.; Hopkins, B. L.; and Hursh, D. E., "The Effects of Study Questions and Grades on Student Test Performance in a College Course," *Journal of Applied Behavior Analysis,* 6: 631-643, 1973.
8. Sherman, J. G., *PSI: Personalized System of Instruction,* W. A. Benjamin, Inc., Menlo Park, Calif., 1974.

THE TRAINING OF PRESERVICE ELEMENTARY SCHOOL TEACHERS IN THE PROCESSES OF SCIENCE

PAUL R. WIDICK
West Chester State College

ABSTRACT

This study was undertaken to examine the relationship between differential science process treatment and the ability of preservice elementary teachers in the performance and application of specific science process skills which were identified for purposes of this study. Seventy-five preservice elementary teachers were randomly assigned to three groups. Two groups were randomly classified as treatment groups, while the third group was employed as a control. Data were collected by post-test only at the end of the 16-week experimental period. The group receiving the integrated process treatment achieved higher scores on the post-test than did the control group. There was no significant difference between the discrete process treatment group and the control group relative to the application of the science process skills.

THE SCIENCE CURRICULUM developments originating during the 1950s have produced changes in the science education of prospective elementary school teachers. These modifications have most commonly been concerned with the process attribute of science.

For the past several decades, science educators have been aware of the necessity of providing preservice elementary teachers with experiences in utilization of the processes of science. These processes refer to the particular operations which one employs during the performance of science as a human enterprise. According to Gagne, traditional science courses have been deficient in accomplishing objectives related to the process component of science (4). Science possesses a dual nature in that it is comprised of both product and process. *Product* is defined as the derivative of the scientific enterprise. The *process* component of science refers to the activity or the method by which the knowledge is derived. It is common knowledge that science courses at all levels have traditionally emphasized science product objectives while simultaneously excluding science process objectives.

The National Association of State Directors of Teacher Education and Certification (NASDTEC), in cooperation

with the American Association for the Advancement of Science (AAAS), has recommended that preservice elementary teachers become involved in experiences with the processes of science. Moreover, NASDTEC-AAAS has suggested that college instructors examine various instructional procedures which could enhance the development of preservice teacher competency in utilizing the science process skill operations (8).

This writer perceives a relationship between the utilization of science process skills and the method of scientific inquiry. The connection is implicit in Bruner's definition of inquiry, described as a series of activities and attitudes concerned with the investigation of any kind of task (3).

There is probably considerable agreement that inquiry includes the process skill operations. Forms of inquiry, however, may be somewhat variable as regards the manner of employment of these skills.

Science educators have vigorously supported the exposure of preservice elementary teachers to the science process skills. It is not unexpected that such exposure could involve numerous techniques. Jacobson (6), however, stresses that a science program for preservice elementary teachers which emphasizes discrete process operations may be less than desirable because the student may not understand the interrelationships of operations.

Research dealing with various inquiry approaches has been conducted with preservice elementary teachers in the past. Olstad (9) reported that subjects exposed to the processes of scientific inquiry showed greater understanding of science than did those subjects exposed to science as subject matter. Menzel (7) concluded that preservice elementary teachers exposed to a laboratory approach showed significantly higher achievement on the science processes of measurement and classification than those subjects exposed to traditional instructional techniques.

The question of teacher competence in the processes of science is in need of further study, and the results of courses and experiences need to be ascertained. Blosser and Howe (2) state that science educators have directed their research efforts more to the training of secondary school teachers rather than elementary school teachers.

Therefore, this study was undertaken to investigate the relationship between differential science process treatment and the competence of preservice teachers in the performance and application of science process operations. Differential science process treatment relative to this study refers to divergent methods of involving preservice elementary teachers with the science process operations.

The null hypotheses proposed for investigation were stated as follows:

1. There is no difference in the performance of specific science process tasks by preservice elementary teachers as a result of differential science process treatment measured by the *Process Instrument for Teachers of Science* (10).

2. There is no difference in the application of science process operations by preservice elementary teachers as a result of differential science process treatment measured by the *Measurement of the Application of Scientific Methodology* (10).

The significance level for testing the null hypotheses was established at 0.05.

For purposes of this study, eleven process skill operations were employed. These skills have been identified by the AAAS as applicable for the science education of preservice elementary school teachers. The eleven skills are: observing; inferring; measuring; predicting; communicating; classifying; defining operationally; formulating hypotheses; analysis of data; interpretation of data; and controlling variables. Several of these skills have been arbitrarily selected and defined as follows:[1]

1. *Inferring*—The process skill which involves the formulation of immediate explanations or conclusions based on prior observations

2. *Communicating*—The process skill involved in the conveyance of an idea by using spoken and/or written words, diagrams, graphs, and other visual aids

3. *Defining operationally*—The process skill involving the definition of terms within the framework of experience

4. *Classifying*—The process skill which involves the ordering of a collection of objects or events

5. *Controlling variables*—The process employed in investigating situations where a number of conditions require uniformity

The differential process treatment identified previously is described as two methods of science process exposure, defined as follows:

1. *Small Increment Process Exposure* (SIPE)— An experimental group exposed to science process treatment utilizing discrete science process operations. *Ss* are involved with single activities involving one of the eleven process skills listed above. A specific or discrete skill is the immediate objective of the activity.

2. *Integrated Process Exposure* (IPE)—An experimental group exposed to science process treatment in which several process skills are employed collectively in the investigation of broad science phenomena. The focus was concerned with the relationship between various skills and the application within the broad investigative framework.

Method

Subjects

During the course registration period, those elementary education majors registering for the required elementary science methods course at West Chester State College were assigned to eight sections. Three of the eight sections were designated for the experiment. Students were randomly as-

signed to all eight sections by the experimenter. Class lists were maintained for all sections, and these lists were carefully scrutinized at the onset of the experiment so that only those students on the class lists were allowed entry into the experimental sections.

The particular type of treatment was arbitrarily determined as each section arrived for the first class meeting. The first group was designated SIPE, the second group IPE, and the third group became the control. No one was informed of the fact that this was an experimental situation.

Experimental Controls

All groups were exposed to the same instructional procedure. The product vehicle was as representative as possible, with the manipulation of the treatment variable occurring through activity selection.

The composition of all treatment groups was determined by random assignment. Moreover, according to the data obtained as a result of a student survey and analyzed by employing a χ^2 test (Table 1), the existence of group homogeneity, from the standpoint of background, can be assumed.

The hours of ten to twelve o'clock were selected as the time for class meetings for all groups. This scheduling was purposely established because of possible extraneous effects should some classes meet in the morning and others in the afternoon. Furthermore, teacher fatigue could have produced another effect should more than one section have met on any one day. All classes for all groups were conducted throughout the experimental period by the same professor, and all classes were held in the same room.

Course requirements were the same for all groups. Students were expected to be in attendance since learning was laboratory-oriented and required direct participation. At no time during the experimental period were intermediate process tests administered to any of the three groups.

The science product as it was defined previously was similar for all treatment groups. The only possible difference was that the product served as the process vehicle for the SIPE and IPE groups, but it was the main objective of the activities experienced by the control group.

During the first class meeting, a survey was administered to determine group equivalence in terms of sex, age, college experience, and other characteristics. Responses were analyzed by use of the χ^2 test (Table 1). Each category identified in Table 1 was subdivided into various classes, which are illustrated as follows:

1. Sex (M and F) ($R = 2$) ($C = 3$) $df = 2$
2. Age (less than 20; between 20-22; between 23-25; greater than 25) ($R = 4$) ($C = 3$) $df = 6$
3. Undergraduate Rank (sophomore; junior; senior); ($R = 3$) ($C = 3$) $df = 4$
4. Prior College Experience (all prior education at West Chester State College; transfer from another four-year college; transfer from a junior college) ($R = 3$) ($C = 3$) $df = 4$
5. Science Courses Completed (one-two courses; three courses; four or more courses) ($R = 3$) ($C = 3$) $df = 4$
6. Teaching Interest (K-3; 4-6; 1-6) ($R = 3$) ($C = 3$) $df = 4$

Table 1.—Chi-Square Values Calculated from Responses to the Student Survey by Three Experimental Groups

Category	*df*	χ^2
Sex	2	1.21
Age	6	4.13
Undergraduate rank	4	7.85
Prior college experience	4	2.44
Science courses completed	4	0.61
Teaching interest	4	2.74

The calculated χ^2 values presented in Table 1 were not significant at the 0.01 significance level. Thus, it was concluded that the three groups were equivalent in terms of responses to the categories indicated.

Procedure

Each group met two periods each week during the 16-week interval. Each weekly class meeting was 75 minutes in length.

All groups were exposed to a total of eight broad science investigations of a type generally found in elementary science methods courses of this variety. For the SIPE group, each investigation was subdivided into small process increments. *Ss* could not proceed beyond a discrete process operation. Such increments may have dealt exclusively with observing, measuring, or any one of the eleven process skills indicated previously. The various process activities were not organized in sequence, and the science media for the SIPE group could be drawn at any time from several of the eight investigations. The IPE *Ss* completed the investigations in an orderly as well as sequential manner. These *Ss* were always aware of the total sequence within each investigation. The IPE group concentrated on a particular broad investigation during any given time. The control group (GE for General Exposure) utilized the same science media except that these *Ss* focused on the science product. This group was concerned with learning science concepts as opposed to science process operations. At no time was the control group intentionally exposed to science process skills.

The two instruments employed in this study for purposes of data collection were the Process Instrument for Teachers of Science and the Measurement of the Application of Scientific Methodology (10). Both instruments were administered as post-tests only at the end of the 16-week experimental period. No tests of any kind were administered during the experiment.

Instruments

The Process Instrument for Teachers of Science is a modified version of the AAAS instrument, Science Process Measure for Teachers-Form B (1). The latter instrument is concerned with the performance of discrete process tasks, i.e., observing, inferring, and the like. The modified version excluded items dealing with behavioral objectives and science process hierarchy relative to the elementary science curriculum project entitled *Science: A Process Approach* (10).

The modified version was administered in 1972 to a randomly selected group of preservice elementary teachers at West Chester State College during a pilot project. The pre-test group was comprised of 23 students, and the post-test group contained 21 of the original 23 participants. No member of this group was involved in any section of the required methods course in science during the semester of the present study. The pre- /post-test results were analyzed by employing the Pearson *r*, which is a correlation based on the mean deviation. The value of the Pearson *r* (0.78) indicated a pre- /post-test correlation significant at the 0.01 level (10).

The Measurement of the Application of Scientific Methodology was designed by the author due to the absence of existing instruments. This measure deals with the application of process operations to selected inquiry situations and presents a written presentation of a scientific activity to which *Ss* respond by agreement, disagreement, or partial agreement to several comments following each situation. The following is presented as an example of such a situation.

SITUATION 2

Another group of students performed the same pendulum experiment and collected the following data from observations taken for ten trials.

Below is the list of observations for the ten trials with the length of string being held constant. Each time measurement is for one single period or one complete movement from the starting position, across, and return.

Length	*Time/Trial (seconds)*									
(cm)	1	2	3	4	5	6	7	8	9	10
?	1.94	1.95	1.93	1.92	1.94	1.96	1.95	1.93	1.92	1.90

The following comments refer to the data collected by the second group of students (Situation 2). Respond to each comment by indicating Agreement (A), Disagreement (D), or Partial Agreement (PA). Indicate by marking the proper column on your answer sheet.

1. The motion of the pendulum for each of the ten trials is uniform in terms of distance covered per time interval.
2. The data should be accepted because of conditions of uniformity which can be assumed for all stials.
3. The conditions under which this data was collected cannot be inferred.
4. The above results for the ten trials are to be expected where experimental controls are enforced.
5. The data for the above ten trials most likely are the result of measuring the time for one single period with a stopwatch rather than calculating the average value based on several trials.
6. The data for the above ten trials most likely are the result of taking the average value of several periods for each trial.

When the Measurement of the Application of Scientific Methodology was administered to the random group of preservice teachers during the 1972 pilot project, pre- /post-test results analyzed by the Pearson *r* indicated a correlation of 0.83 which was judged as significant at the 0.01 level (10).

Results

The scores achieved by all groups on the post-tests were subjected to analysis employing a one-way analysis of variance and the Scheffé test of multiple comparisons. Tables 2 and 3 provide statistical data for all groups relative to post-test scores.

Table 2.—Post-Test Results for the *Process Instrument for Teachers of Science*

Group	*N*	*Mean*	*Variance*	*SD*
SIPE	20	27.95	33.94	5.83
IPE	20	27.55	42.74	6.04
GE	20	23.25	24.41	4.94

The significance level established by the experimenter for testing the null hypotheses was 0.05.

Table 3.—Post-Test Results for the *Measurement of the Application of Scientific Methodology*

Group	*N*	*Mean*	*Variance*	*SD*
SIPE	20	33.75	9.67	3.11
IPE	20	35.20	12.27	3.50
GE	20	31.90	22.56	4.75

The significance level established by the experimenter for testing the null hypotheses was 0.05.

Tables 4 and 5 provide analysis of experimental results based on a one-way analysis of variance.

For significance at the 0.05 level, the calculated *F*-value for 2 and 57 degrees of freedom must equal or exceed the tabled value of 3.16. Therefore, the calculated *F* (4.296) is significant at the 0.05 level, and the null hypothesis of no significant difference between the three groups measured by the process instrument was rejected.

For significance at the 0.05 level, the calculated *F*-value for 2 and 57 degrees of freedom must equal or exceed 3.16. The calculated *F* (3.692) is significant at the 0.05 level. Therefore, the null hypothesis of no significant difference between the three groups as measured by the applications instrument was rejected.

Table 4.—ANOVA Results for the *Process Instrument for Teachers of Science*

Source	*SS*	*df*	*MS*	*F*
Between	271.60	2	135.80	4.296*
Within	1801.65	57	31.61	
Total	2073.25	59		

* Significant at the 0.05 level

Table 5.—ANOVA Results for the *Measurement of the Application of Scientific Methodology*

Source	*SS*	*df*	*MS*	*F*
Between	109.44	2	54.72	3.692*
Within	844.75	57	14.82	
Total	954.19	59		

* Significant at the 0.05 level

The Scheffé test of multiple comparisons (*S*-method) was employed to determine differences between group means for the post-tests. The Scheffé test can be employed to test the significance of difference between means separately for all pairs of means where multiple groups are employed. The *S*-method is defined as follows:

$$\Psi = c_1 \overline{X}_1 + c_2 \overline{X}_2 + c_3 \overline{X}_3$$

where:

the constants c_1, c_2, c_3 are positive and negative real numbers that sum to 0, and

Ψ is the estimate of the contrast between means

Before the significance of any contrast Ψ can be judged, the variance of the contrast must be determined. The variance can be estimated by σ^2_{ψ}

$$\sigma^2_{\psi} = MS_w \left(\frac{c_1^2}{N_1} + \frac{c_2^2}{N_2} + \frac{c_3^2}{N_3} \right)$$

where:

σ^2_{ψ} is the estimated variance of the contrast Ψ, and

MS_w is the mean square within groups

The hypothesis that $\Psi = 0$ is rejected if the absolute value of the ratio exceeds the square root of $(J - 1)$ times the percentile point. Reject the null hypothesis that $\Psi = 0$ if

$$\frac{\Psi}{\sigma_{\psi}} > \sqrt{(J-1)_{.95\,F\,(J-1)\,(N-J)}}$$

The value of $\sqrt{(J-1)_{.95\,F\,(2.57)}}$ is $\sqrt{2\,(3.16)}$ or 2.514.

Therefore, the null hypothesis of no significant difference can be rejected when the *F*-ratio $\left(\frac{\Psi}{\sigma_{\psi}}\right)$ exceeds 2.514.

Tables 6 and 7 present the comparisons between group means on the post-tests by analysis employing the *S*-method.

The post-test results for the process instrument indicated significant differences for the three experimental groups (Table 6). There was no significant difference between the means of the SIPE and IPE groups. The achievement by the SIPE and IPE groups, however, differed significantly from the achievement by the GE control group.

The mean scores for the SIPE and IPE groups were not significantly different. This outcome should not be unexpected in view of the fact that both groups experienced direct science process treatment using science process skills. The difference involved the method of exposure.

The analysis of group means on the applications instrument (Table 7) indicated no significant difference between the SIPE and GE groups. A significant difference existed between the means of the IPE and GE groups.

Although the mean score achieved by the SIPE group on the Measurement of the Application of Scientific Methodology was higher than that achieved by the GE group, the results were not statistically different at the 0.05 significance level. This outcome could possibly indicate that the exposure to singular, discrete process skill activities does not have any significant effects in terms of the transfer of these skills relative to functioning within a broader inquiry framework. Apparently, the SIPE group could not utilize the process skills in inquiry situations in a manner superior to the GE group despite the fact that the SIPE group had purposely been exposed to discrete process skill experiences and the control group had received no such exposure.

The mean score achieved by the IPE group on the Measurement of the Application of Scientific Methodology was statistically different from the mean achieved by the control group. This result could indicate that the achievement of the IPE group was superior to that of the GE group because of the integrated process skill treatment where the process skills were emphasized in an interrelated fashion.

Discussion

The development of competency in science processes when examined within the framework of performance and application, as measured by the Process Instrument for Teachers of Science and the Measurement of the Application of Scientific Methodology, does appear to be influenced by the type of process exposure in science for preservice elementary school teachers. The experimental group exposed to integrated process skill activities appeared to demonstrate superiority over the other groups in the application of these skills to broader inquiry situations. These results tend to verify the concerns voiced by Jacobson (6)

Table 6.–Scheffé Test Analysis for the Process Instrument

Contrast		Ψ	σ^2_Ψ	Ψ/σ_Ψ	F
$\bar{X}$ (SIPE)	$\bar{X}$ (IPE)	.40	3.16	.40/1.777	0.225
$\bar{X}$ (SIPE)	$\bar{X}$ (GE)	5.05	3.16	5.05/1.777	2.842*
$\bar{X}$ (IPE)	$\bar{X}$ (GE)	4.65	3.16	4.65/1.777	2.617*

*Significant at the 0.05 level

Table 7.–Scheffé Test Analysis for the Applications Instrument

Contrast		Ψ	σ^2_Ψ	Ψ/σ_Ψ	F
$\bar{X}$ (SIPE)	$\bar{X}$ (IPE)	−1.45	1.48	−1.45/1.22	−1.189
$\bar{X}$ (SIPE)	$\bar{X}$ (GE)	1.85	1.48	1.85/1.22	1.516
$\bar{X}$ (IPE)	$\bar{X}$ (GE)	−3.30	1.48	−3.30/1.22	2.705*

*Significant at the 0.05 level

in that a science program for preservice elementary teachers which emphasizes the discrete process skills may not be desirable because the student may not understand the interrelationships of process operations in science.

In truly scientific inquiry, process skills are utilized as tools and integrated within a broad context, enabling the experimenter to formulate decisions and generalizations concerning descriptions and explanations of scientific phenomena. The application of a single skill or operation does not follow automatically because one has been exposed to that skill. The skill or operation must be experienced within a somewhat realistic domain which might suggest and even facilitate the transfer of that skill to broader-based inquiry situations.

Implications

The process of scientific inquiry, which may be described as mental skills and habits essential to rational thinking, has historically been given high priority as an objective of elementary education. Unfortunately, this objective has been most commonly a subject of discourse rather than practice.

In order for the development of rational thinking to become an educational reality, inquiry practices in teacher education must be subjected to intensive analyses. The following recommendations are suggested for future study:

1. Competency levels for preservice elementary teachers as regards inquiry skills should be identified.
2. Identification of various instructional procedures for inquiry training are in order. Included in this recommendation are content variations in science.
3. Research dealing with various methods of teaching inquiry should be undertaken.
4. Development of suitable instruments to measure competency of preservice teachers requires considerable effort.

FOOTNOTE

1. See Commission on Science Education, *Science: A Process Approach–Commentary for Teachers*, American Association for the Advancement of Science/Xerox Corporation, 1970.

REFERENCES

1. American Association for the Advancement of Science, Commission on Science Education, *Science Process Measure for Teachers*, AAAS Miscellaneous Publication 69-9, 1969.
2. Blosser, P.E.; and Howe, R.W., "An Analysis of Research on Elementary Teacher Education Related to the Teaching of Science," *Science and Children*, 6:50-60, January-February 1969.
3. Bruner, J.S., "The Act of Discovery," *Harvard Educational Review*, 31:21-32, No. 1, 1966.
4. Gagne, R.M., "The Learning Requirements for Enquiry," *Journal of Research in Science Teaching*, 1:144-153, No. 2, 1966.
5. Glass, G.V.; and Stanley, J.C., *Statistical Methods in Education and Psychology*, Prentice Hall, Englewood Cliffs, N.J., 1970.
6. Jacobson, W.J., "Teacher Education and Elementary School Science–1980," *The Journal of Research in Science Teaching*, 5:73-80, No. 1, 1968.
7. Menzel, E.W., "A Study of Preservice Elementary Teacher Education in Two Processes of Science," unpublished doctoral dissertation, Temple University, 1968.
8. National Association of State Directors of Teacher Education and Certification–American Association for the Advancement of Science, *Guidelines for Science and Mathematics in the Preparation of Elementary School Teachers*, AAAS Miscellaneous Publication 63-7, 1963.
9. Olstad, R.G., "The Effect of Science Teaching Methods on the Understanding of Science," *Science Education*, 53:9-11, February 1969.
10. Widick, P.R., "An Analysis of Three Methods of Instruction Employed with Preservice Elementary School Teachers to Develop Competency in Science Process Skills," unpublished doctoral dissertation, Temple University, copyright 1973.

DIRECTIONS FOR J.E.E. CONTRIBUTORS

The Journal of Experimental Education publishes specialized or technical education studies, treatises about the mathematics or methodology of behavioral research, and monographs of major current research interest.

ABSTRACT REQUIRED

An abstract of not more than 120 words must accompany each manuscript. It should precede the text, be designated *ABSTRACT*, and conform to the following standards:

1. In a research paper, include statements of (a) the problem, (b) the method, (c) the data, and (d) the conclusions.

2. In a review or discussion article, state the topics covered and the central thesis.

3. Standard abbreviations may be used freely if the meaning is clear. Use complete sentences and do not repeat information which is in the title.

TEXT

Usually research articles should have a clearly organized order of presentation. The following elements and their order is a guide and is not intended to be prescriptive.

The Problem. The nature, scope, and significance of the problem should be presented.

Related Research. Presentation and discussion of related research should be selective and should include references essential to clarify the problem under examination.

Methodology. This section should consist of hypotheses, description of the sample and sampling procedures, discussion of the variables, description of the data-gathering instruments (if well-known, their description may be omitted), and general description of the statistical procedures.

Presentation and Analysis of Data. Analysis of the data and conclusions about the hypotheses should be more than mere presentation. Rival hypotheses and significance for related studies and educational theory should be considered. Summary data (means, standard deviations, frequencies, correlation coefficients, etc.) should be presented in tabular or graphic form. Discussion of these data should be interpretive.

Summarizing Statements. A summary of conclusions and implications for education may supplement the abstract.

STYLE

Certain suggestions are offered here to achieve uniformity in publication style and to conserve the time and energy of the author and editorial staff in the preparation of copy. *A Manual of Style,* 12th ed., University of Chicago Press, Chicago, 1960, may be used as a style manual in preparation of manuscripts.

Two Copies Required. Copies should be double spaced with wide margins. Typed copies are preferred, but dittoed or mimeographed copies will be accepted if they are legible.

Subheads. Articles are frequently improved by the judicious use of subheads. However, avoid use of the title, INTRODUCTION, for a lead section.

Title. Try to use a short title, preferably no more than ten words. Avoid superfluous phrases, such as "A Comparison of . . ." "A Study of . . .," and "The Effectiveness of . . . ."

Tables. Prepare tables precisely as they are to appear in *The Journal,* caption each with a brief and to-the-point title, and number consecutively with arabic numerals: *Table 3.* INTERCORRELATION MATRIX, VARIABLES OPTIMALLY ORDERED.

Figures. Draw any graphs and charts in black ink on good quality paper, title each, and number consecutively, thus: *Figure 4.* SCHOOL ENROLLMENT. Send the original copy. We cannot accept mimeographed figures or charts. Data should not be framed, and ordinates and abscissas should be shortened to occupy as little space as possible.

Tables and Figures. Tables and figures must be original copies acceptable for reproduction. A charge will be assessed for any redrawing or re-typing of tables or figures.

Technical Symbols. All symbols, equations, and formulas must be clearly represented. Differentiate between numbers and letters (zero and the letter o; one and the letter l, etc.) Identify hand-drawn Greek letters in the margin. Align subscripts carefully.

Footnotes. Avoid explanatory footnotes by incorporating their content in the text. However, for essential footnotes, identify them with consecutive superscripts, as *book,*[2] *study,*[3] etc., and list the footnotes in a section, entitled FOOTNOTES, at the end of the text, but preceding the REFERENCES.

References. References should be listed alphabetically according to the author's last name at the end of the manuscript. Each entry should be numbered. Citation of a reference in the text should be made with this number, as (2); or if specific pages are to be indicated, as (2:245–7).

Illustrations of article and book references:

1. Bittner, Reign H. and Wilder, Carlton E., "Expectancy Tables: A Method of Interpreting Correlation Coefficients," *The Journal of Experimental Education,* 14:245–52, March 1946.

2. Thomson, Godfrey, H., *The Factorial Analysis of Human Ability,* Houghton Mifflin Co., Boston, 1950, 383 pp.

PROCEDURES

Send manuscripts to John Schmid, Department of Research and Statistical Methodology, University of Northern Colorado, Greeley, CO 80631.

Each contributor will receive 2 complimentary copies of the issue in which his article appears. Keep an exact carbon copy of your manuscript so that if a question arises the editors can refer you to specific pages, paragraphs, or lines for clarification by letter.

THE JOURNAL OF EXPERIMENTAL EDUCATION

4000 Albemarle Street, N.W., Suite 302,
Washington, D.C. 20016

Return Postage Guaranteed

Second Class Postage Paid at Washington, D.C.